Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Exploração de Modelos Híbridos com Atenção para Classificação de Ceratite Infecciosa em Imagens de Microscopia Confocal in Vivo

Isaac Silva Santos Ramos

Teresina-PI, Fevereiro de 2026

Isaac Silva Santos Ramos

Exploração de Modelos Híbridos com Atenção para Classificação de Ceratite Infecciosa em Imagens de Microscopia Confocal in Vivo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Ivan Saraiva Silva

Coorientador: Vinicius Ponte Machado

Teresina-PI

Fevereiro de 2026

Ficha elaborada de acordo com o Código de Catalogação Anglo-Americano AACR2

R175e

Ramos, Isaac Silva Santos.
Exploração de modelos híbridos com atenção para classificação de ceratite infecciosa em imagens de microscopia confocal in vivo. / Isaac Silva Santos Ramos. – Teresina, PI, 2026.
69 f.: il. (alg. color.)

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Programa de Pós-Graduação em Ciência da Computação, 2026.

Orientador: Prof. Dr. Ivan Saraiva Silva.
Coorientador: Prof. Dr. Vinicius Ponte Machado.

1. Redes neurais (Computação). 2. Ceratite. 3. Córnea - Doenças. 4. Aprendizado do computador. I. Silva, Ivan Saraiva. II. Machado, Vinicius Ponte. III. Título.

CDU: 004.8:617.7
CDD: 006.31

Ficha elaborada pelo Serviço de Catalogação da Biblioteca Prof. Queirois, IFMA, Campus Caxias
Ianna Torres Lustosa (Bibliotecária) CRB-13/687

Isaac Silva Santos Ramos

Exploração de Modelos Híbridos com Atenção para Classificação de Ceratite Infecciosa em Imagens de Microscopia Confocal in Vivo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Documento assinado digitalmente
gov.br
IVAN SARAIVA SILVA
Data: 17/03/2026 11:33:30-0300
Verifique em https://validar.iti.gov.br

Ivan Saraiva Silva
Orientador

Documento assinado digitalmente
gov.br
VINICIUS PONTE MACHADO
Data: 18/03/2026 15:02:06-0300
Verifique em https://validar.iti.gov.br

Vinicius Ponte Machado
Co-Orientador

Documento assinado digitalmente
gov.br
RODRIGO DE MELO SOUZA VERAS
Data: 19/03/2026 08:47:51-0300
Verifique em https://validar.iti.gov.br

Rodrigo de Melo Souza Veras
Convidado 1

Documento assinado digitalmente
gov.br
KELSON ROMULO TEIXEIRA AIRES
Data: 18/03/2026 17:10:04-0300
Verifique em https://validar.iti.gov.br

Kelson Romulo Teixeira Aires
Convidado 2

Documento assinado digitalmente
gov.br
ANNE MAGALY DE PAULA CANUTO
Data: 20/03/2026 10:37:00-0300
Verifique em https://validar.iti.gov.br

Anne Magaly de Paula Canuto
Convidado 3

Teresina-PI
Fevereiro de 2026

Agradecimentos

Agradeço a Deus.

Sou grato aos meus pais, Romualdo e Gizelia, por ajudarem nos momentos mais difíceis. Agradeço por toda a dedicação e carinho que tiveram durante as fases da minha vida. Sem eles eu não sou quem sou hoje.

Agradeço ao meu orientador, Prof. Ivan Saraiva, por todos os conselhos, ideias, pela paciência e ajuda nesse período. Um orientador que sabe lidar nos momentos difíceis, de reformulação da pesquisa, que me ouviu várias veze e que sempre torce pelo melhor. Obrigado ao co-orientador, Prof. Vinicius Machado, que temos uma parceria e sempre agradeço por isso. Aos professores Kelson Aires e Rodrigo Veras, por expor suas observações do trabalho durante a qualificação, visando sempre a melhor forma de contribuição científica.

Aos meus amigos que, indiretamente, contribuíram de alguma forma em superar meus limites, não desistindo em nenhum momento. Com os amigos, consigo desabafar e seguir em frente, sempre em frente.

"A atenção é a mais importante
de todas as faculdades para
o desenvolvimento da
inteligência humana."
(Charles Darwin)

Resumo

A ceratite infecciosa (CI) é uma inflamação da córnea causada por agentes patológicos, como fungos e protozoários do gênero *Acanthamoeba*, podendo evoluir para cegueira quando não diagnosticada precocemente. Embora a cultura microbiológica seja um método preciso de diagnóstico, sua natureza invasiva e o tempo de resposta limitam o uso clínico. Nesse contexto, a microscopia confocal *in vivo* (IVCM) surge como alternativa não invasiva e de alta resolução, porém sua interpretação demanda expertise especializada. Este trabalho investiga modelos de *deep learning* para a classificação automática de imagens IVCM em quatro classes: ceratite fúngica (FK), ceratite por *Acanthamoeba* (AK), ceratite não especificada (NSK) e córnea normal. A partir da base pública IVCM-Keratitis, foram avaliadas quatro arquiteturas convolucionais individuais (VGG16, MobileNetV3, DenseNet161 e ResNet101) e quatro abordagens, baseadas na concatenação de vetores de características de duas arquiteturas base. As abordagens mais completas incorporam um bloco de atenção do tipo *Self-Attention* (SA) após a concatenação, permitindo ponderar a relevância relativa das características antes da classificação por uma MLP. Os resultados indicam que a Abordagem IV (DenseNet161 + ResNet101 com SA) apresentou o melhor equilíbrio entre detecção e erros de classificação, alcançando médias de precisão de 96,22%, sensibilidade de 96,01% e *F1-score* de 96,03% (acurácia média de 96,01%). Por fim, aplicou-se o algoritmo Grad-CAM para interpretar qualitativamente as regiões mais relevantes nas predições, evidenciando padrões de atenção complementares entre as arquiteturas base. Os achados sugerem que modelos híbridos com SA podem contribuir para diagnósticos assistidos em contextos com escassez de especialistas, apoiando a tomada de decisão clínica e reduzindo casos de perda visual evitável.

Palavras-chave: Microscopia Confocal *in vivo*. Ceratite infecciosa. Aprendizado profundo. Modelos híbridos. *Self-Attention*. Interpretabilidade (Grad-CAM).

Abstract

Infectious keratitis (IK) is an inflammation of the cornea caused by pathogenic agents such as fungi and protozoa of the genus Acanthamoeba, and it may progress to blindness when not diagnosed early. Although microbiological culture is an accurate diagnostic method, its invasive nature and turnaround time limit routine clinical use. In this context, in vivo confocal microscopy (IVCM) emerges as a non-invasive, high-resolution alternative; however, its interpretation requires specialized expertise. This study investigates deep learning models for automatic classification of IVCM images into four categories: fungal keratitis (FK), Acanthamoeba keratitis (AK), non-specified keratitis (NSK), and normal cornea. Using the public IVCM-Keratitis dataset, we evaluated four individual convolutional architectures (VGG16, MobileNetV3, DenseNet161, and ResNet101) and four hybrid approaches based on concatenating feature vectors extracted by two backbone networks. The most comprehensive variants incorporate a self-attention (SA) block after concatenation, enabling adaptive weighting of the relative importance of features prior to classification by an MLP. The results indicate that Approach IV (DenseNet161 + ResNet101 with SA) achieved the best balance between detection performance and misclassification errors, with mean precision of 96.22%, sensitivity of 96.01%, and F1-score of 96.03% (mean accuracy of 96.01%). Finally, Grad-CAM was employed to qualitatively interpret the most relevant regions driving the predictions, revealing complementary attention patterns between the backbone networks. Overall, the findings suggest that SA-enabled hybrid models can support computer-aided diagnosis in settings with limited specialist availability, assisting clinical decision-making and reducing avoidable vision loss.

Keywords: In vivo confocal microscopy. Infectious keratitis. Deep learning. Hybrid models. Self-attention. Explainable AI (Grad-CAM).

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CI	Ceratite Infecciosa.
IA	Inteligência Artificial.
IVCM	*in vivo Confocal Microscopy*.
MLP	*Multilayer Perceptron*.
CNN	*Convolutional Neural Networks*.
KOH	Hidróxido de potássio.
PCR	*Polymerase Chain Reaction*.
ML	*Machine Learning*.
AK	*Acanthamoeba Keratitis - Ceratite por Acanthamoeba*.
FK	*Fungal Keratitis - Ceratite Fúngica*.
NSK	*Non-Specified Keratitis - Ceratite Não-Especificada*.
ROC	Curva de Característica de Operação do Receptor.
HMF	*Histogram Matching Frequency*.
XAI	*Explainable Artificial Intelligence*.
JPEG	*Joint Photographic Experts Group*.
ReLU	*Rectified Linear Unit*.
NAS	*Neural Architecture Search*.
SE	*Squeeze-and-Excitation*.
SA	*Self-Attention*
MHSA	*Multi-head Self-attention*.
Grad-CAM	*Gradient-weighted Class Activation Mapping*.
DL	*Deep Learning*.
AUC	*Area Under the Curve*.

ViT	*Vision Transformer*.
FLOPs	*Floating Point Operations per Second*.
MB	*Megabytes*.
BN	*Batch Normalization*.
SCS	*Subarea Contrast Stretching*.

Lista de símbolos

$\in$	Pertence - utilizado para denotar pertencimento a um conjunto.
Σ	Somatório - utilizado para denotar uma soma de elementos.
$\cdot$	Produto escalar - utilizado para denotar uma multiplicação.
$\odot$	Produto de Hadamard - utilizado para denotar uma multiplicação ponto a ponto.
$A^\top$	Transposta de matriz - utilizado para denotar uma matriz transposta.
α, β	Vetores de atenção - utilizado para denotar vetores de atenção.
$\times$	Dimensão - utilizado para denotar o tamanho de uma imagem.
$+$	Soma - utilizado para denotar soma simples.
$\oplus$	Soma vetorial - utilizado para denotar soma entre vetores.

Sumário

1 Introdução

A Ceratite Infecciosa (CI) é uma importante causa de perda visual e cegueira em nível global, configurando-se como um desafio relevante para a saúde pública e a prática oftalmológica (AUSTIN; LIETMAN; ROSE-NUSSBAUMER, 2017). Trata-se de um processo inflamatório da córnea desencadeado por diferentes agentes etiológicos, incluindo fungos (Ceratite Fúngica), bactérias (Ceratite Bacteriana), vírus (Ceratite Viral) e protozoários. Em todos os casos, a doença pode evoluir rapidamente e o atraso no diagnóstico e no início do tratamento aumenta substancialmente o risco de sequelas permanentes e cegueira (ESSALAT et al., 2023).

A distribuição da CI apresenta forte variação geográfica, refletindo disparidades socioeconômicas, condições ambientais e acesso a serviços de saúde. Em países de alta renda, como Estados Unidos, Inglaterra e Austrália, as taxas de incidência relatadas foram de 27,6, 40,3 e 6,6 casos por 100.000 habitantes nos anos de 1999, 2006 e 2015, respectivamente (CABRERA-AGUAS; KHOO; WATSON, 2022). Em contraste, em países de baixa e média renda, a carga da doença é substancialmente maior: na Índia, a taxa estimada alcança 113 casos por 100.000 habitantes, enquanto no Nepal atinge 799 casos por 100.000 habitantes (CABRERA-AGUAS; KHOO; WATSON, 2022). Esses números evidenciam a magnitude do problema em contextos com maior vulnerabilidade social.

No Brasil, entre 2018 e 2022, foi registrada uma média anual de 5.475 casos confirmados de CI, com pico de notificações em 2019 (6.290 casos) Adriano et al. (2024). A taxa média nacional no período foi de 2,48 casos por 100.000 habitantes, com predomínio de casos no sexo feminino (54%). A faixa etária mais acometida correspondeu a adultos entre 20 e 59 anos, totalizando 13.625 diagnósticos (ADRIANO et al., 2024). Esses dados reforçam que a CI afeta majoritariamente indivíduos em idade produtiva, com implicações sociais e econômicas relevantes.

O diagnóstico convencional da CI baseia-se predominantemente na avaliação clínica, complementada por exames laboratoriais. A cultura microbiológica do raspado da córnea permanece como o padrão-ouro para a identificação do agente etiológico. Este procedimento é, entretanto, invasivo, tecnicamente exigente e demanda um tempo médio de aproximadamente 72 horas para a emissão do laudo (SILVA et al., 2025). Em cenários de urgência e de recursos limitados, essas características dificultam a tomada de decisão terapêutica precoce, podendo comprometer o prognóstico visual.

Nesse contexto, a microscopia confocal *in vivo* (IVCM) da córnea surge como uma alternativa não invasiva, capaz de capturar imagens seccionais de alta resolução de todas as camadas corneanas, por meio de um *scanner* a laser, com maior agilidade diagnóstica

(KRYSZAN et al., 2024). Essa técnica permite visualizar microestruturas compatíveis com a presença de agentes infecciosos, como cistos de *Acanthamoeba* e hifas fúngicas, contribuindo para o diagnóstico etiológico em tempo reduzido.

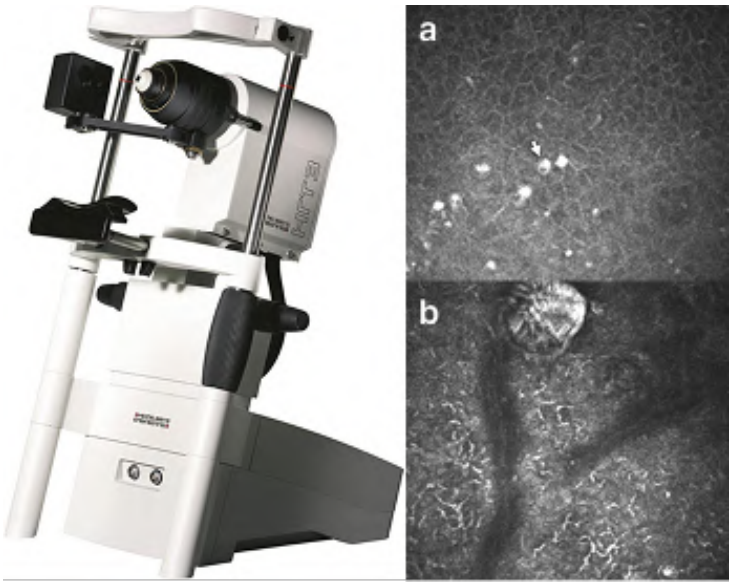


Figura 1 – Sistema de microscopia confocal *in vivo* com *scanner* a laser (à esquerda) e exemplos de imagens seccionais da córnea obtidas *in vivo* (à direita): (a) presença de estruturas hiper-refletivas no estroma; (b) exemplo de estrutura de maior dimensão e alta refletividade observada em imagem IVCM.

Apesar de suas vantagens, a interpretação de imagens IVCM é altamente dependente de especialistas experientes e pode ser afetada pela variabilidade interobservador. Além disso, o volume crescente de exames e a escassez de profissionais especializados tornam o processo de análise visual manual potencialmente demorado e sujeito a inconsistências. Esse cenário motiva o uso de métodos computacionais que auxiliem a padronizar e acelerar a interpretação das imagens.

A área de visão computacional, em conjunto com técnicas de aprendizado profundo (*deep learning*), tem se consolidado como uma abordagem promissora para a análise automatizada de imagens médicas, incluindo tarefas de segmentação e classificação (VAZQUEZ, 2024). Redes Neurais Convolucionais (CNNs) permitem extrair automaticamente representações discriminativas a partir dos dados de imagem, superando limitações de abordagens baseadas em descritores manuais e em etapas extensas de pré-processamento. No contexto da oftalmologia, o uso de CNNs tem se expandido para o estudo de diversas estruturas do segmento anterior e posterior do olho, explorando imagens provenientes de diferentes modalidades de aquisição.

Segundo Kryszan et al. (2024), o desenvolvimento de sistemas de aprendizado profundo voltados especificamente para a classificação do segmento anterior do olho, a partir de imagens IVCM, ainda enfrenta inúmeros desafios. Entre eles, destacam-se o número limitado de bases de dados anotadas, a heterogeneidade morfológica das estru-

turas observadas, a presença de ruídos e artefatos de aquisição, bem como a necessidade de modelos mais interpretáveis, capazes de fornecer explicações visuais que auxiliem o especialista na validação do diagnóstico automatizado. A revisão contínua de estudos científicos nessa área é essencial para consolidar o conhecimento e orientar o desenho de novas arquiteturas que conciliem desempenho, robustez e explicabilidade.

Nesse cenário, métodos de visão computacional e aprendizado profundo têm sido explorados para automatizar a análise de imagens de microscopia confocal, com resultados promissores na detecção e classificação de casos de CI. No entanto, ainda há espaço para o desenvolvimento de modelos que explorem de forma mais eficaz a complementaridade entre diferentes extratores de características, bem como para uso de estratégias de explicabilidade que tornem as decisões dos modelos mais interpretáveis.

Este trabalho insere-se nesse contexto ao propor um método de classificação automática de imagens IVCM da base IVCM-Keratitis, fundamentado em arquiteturas híbridas de Redes Neurais Convolucionais. A abordagem combina (i) dois ou mais fluxos convolucionais paralelos para extração de características, (ii) um mecanismo de atenção do tipo *self-attention* aplicado na etapa de fusão, e (iii) uma estratégia de explicabilidade baseada em Grad-CAM. A hipótese central é que a integração de representações aprendidas por backbones com vieses arquiteturais distintos, quando acoplada a um mecanismo de ponderação adaptativa na fusão e a mapas de ativação interpretáveis, pode simultaneamente elevar o desempenho diagnóstico e oferecer suporte visual ao especialista na avaliação de casos de Ceratite Infecciosa (CI).

No escopo deste estudo, o termo "híbrido" refere-se, de forma estrita, a modelos que integram em um único classificador duas (ou mais) arquiteturas convolucionais pré-definidas, as quais processam a mesma imagem de entrada em fluxos paralelos. Cada fluxo aprende descritores em diferentes níveis de abstração e com diferentes indutivos arquiteturais, resultando em mapas de características complementares. Em seguida, esses mapas são alinhados e combinados em uma camada de hibridização, na qual a concatenação é assistida por um módulo de *self-attention*. Esse módulo atua como um mecanismo de recalibração dinâmica, atribuindo pesos relativos às informações provenientes de cada fluxo e, assim, modulando sua contribuição para a decisão final. A Figura 2 apresenta a arquitetura geral proposta, destacando o ponto de fusão dos mapas de características e a aplicação do mecanismo de atenção.

1.1 Motivação

A CI é tratada como uma condição oftalmológica de emergência, na qual o atraso no diagnóstico e no início do tratamento adequado está diretamente associado ao risco de perfuração corneana, necessidade de transplante e cegueira irreversível. Em muitos cenários,

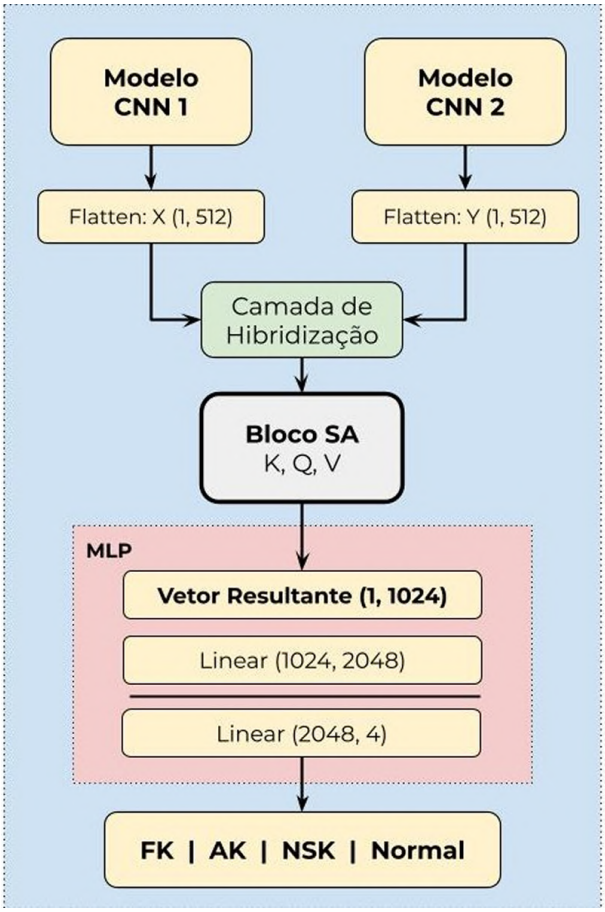


Figura 2 – Visão geral da arquitetura híbrida proposta para classificação de imagens IVCM. Duas redes convolucionais (CNN 1 e CNN 2) extraem representações vetoriais de dimensão 512, que são integradas na camada de hibridização. Em seguida, um bloco SA (Q, K, V) recalibra a contribuição de cada fluxo e produz um vetor combinado (1×1024), utilizado por uma MLP para a predição das quatro classes.

especialmente em serviços com alta demanda e recursos limitados, o acesso a métodos diagnósticos especializados, como a microscopia confocal e a cultura microbiológica, é restrito ou sujeito a longos tempos de espera.

Mesmo quando a microscopia confocal está disponível, a interpretação das imagens pode ser desafiadora, exigindo experiência específica e estando sujeita a variabilidade entre observadores. Paralelamente, as bases públicas de imagens IVCM ainda são relativamente pequenas e heterogêneas, o que dificulta a construção de modelos robustos de aprendizado profundo e impõe desafios adicionais de generalização e explicabilidade.

Diante desse contexto, justifica-se o desenvolvimento de modelos de aprendizado profundo que combinem alto desempenho de classificação, eficiência computacional e mecanismos de explicabilidade, de forma a apoiar o especialista na tomada de decisão. Em particular, espera-se que um sistema capaz de identificar automaticamente padrões compatíveis com ceratite infecciosa em imagens IVCM possa fornecer suporte computaci-

onal para a detecção precoce dessa condição, contribuindo para a redução das taxas de cegueira.

1.2 Objetivos

O objetivo geral deste trabalho é propor e avaliar arquiteturas híbridas de redes neurais convolucionais, associadas a mecanismos de self-attention e técnicas de explicabilidade, para a classificação da ceratite infecciosa em imagens de microscopia confocal *in vivo* da base IVCM-Keratitis.

1.2.1 Objetivos Específicos

Para atingir esse objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- Realizar uma revisão da literatura sobre métodos de aprendizado profundo aplicados à análise de imagens IVCM e à detecção de ceratite infecciosa;
- Organizar e pré-processar a base IVCM-Keratitis, incluindo a remoção de imagens ruidosas e o balanceamento das classes;
- Investigar arquiteturas convolucionais híbridas e a incorporação de blocos de self-attention;
- Comparar o desempenho das arquiteturas propostas com modelos de referência da literatura, por meio de métricas quantitativas de classificação e análise de eficiência computacional;
- Empregar técnicas de explicabilidade, em especial o *Gradient-weighted Class Activation Mapping* (Grad-CAM), para interpretar as decisões do modelo e avaliar a consistência clínica das regiões de atenção.

Supõe-se que, em arquiteturas de redes neurais convolucionais, a combinação de duas redes complementares e um bloco de self-attention apresente desempenho de classificação estatisticamente superior ao de cada rede isolada, mantendo eficiência computacional compatível com o uso em cenários clínicos para detecção de ceratite infecciosa em imagens IVCM-Keratitis.

Em termos metodológicos, este estudo utiliza a base de dados pública IVCM-Keratitis, composta por imagens de microscopia confocal in vivo de córnea classificadas em quatro categorias (*Acanthamoeba*, Fungo, Não Especificado e Córnea Normal). Após a etapa de limpeza das imagens e balanceamento das classes, o conjunto de dados é

particionado em subconjuntos de treinamento, validação e teste, com validação cruzada estratificada em cinco *folds*.

As arquiteturas propostas empregam modelos convolucionais pré-treinados como arquiteturas de extração de características, cujos vetores de saída são combinados por concatenação e processados por um classificador totalmente conectado com bloco de self-attention. O treinamento é realizado com técnicas de aumento de dados e early stopping, e a avaliação inclui métricas como acurácia, F1-score, sensibilidade, especificidade e área sob a curva ROC. Por fim, mapas de calor Grad-CAM são gerados para analisar as regiões de interesse utilizadas pelo modelo na tomada de decisão.

1.3 Organização do Trabalho

Esta dissertação está estruturada em seis capítulos. O Capítulo 1 apresenta o tema da pesquisa, situando o problema da ceratite infecciosa e discutindo a relevância do uso de aprendizado profundo para a classificação de imagens médicas, com ênfase na necessidade de modelos explicáveis, leves e eficazes. Também é delimitado que o estudo propõe um sistema de apoio à decisão baseado em imagens, sem substituir o diagnóstico clínico.

O Capítulo 2 reúne o referencial teórico e os fundamentos técnicos relacionados à anatomia da córnea, às manifestações morfológicas da ceratite infecciosa em imagens IVCM e aos principais conceitos de redes neurais convolucionais (CNNs), atenção e interpretabilidade. O Capítulo 3 revisa os principais trabalhos correlatos e discute as lacunas que motivam a presente proposta.

O Capítulo 4 descreve os materiais utilizados, incluindo a base de dados IVCM-Keratitis, os modelos de CNN utilizados, as técnicas de pré-processamento aplicadas, as métricas de avaliação e o algoritmo Grad-CAM. O Capítulo 5 detalha a metodologia experimental, apresentando as arquiteturas exploradas, que combinam vetores extraídos do VGG16, MobileNetV3, DenseNet161, ResNet101, bloco de atenção e a explicabilidade com o Grad-CAM. Por fim, o Capítulo 6 apresenta as conclusões, as limitações do estudo e as perspectivas para trabalhos futuros, sobretudo no aprimoramento da interpretabilidade e na integração clínica do modelo.

2 Referencial Teórico

A inteligência artificial desempenha um papel relevante no campo da medicina, onde produz impactos significativos (PICCIALLI et al., 2021). Um desses impactos é a capacidade de tomar decisões para classificar imagens, auxiliando no processo de diagnóstico realizado pelo profissional.

Ainda é comum observar-se casos em que o profissional pode estar errado, possibilitando a ocorrência de complicações no tratamento. Diagnósticos incorretos podem decorrer de interpretações imprecisas dos dados, atribuídas à elevada complexidade das imagens. Esses desafios podem ser mitigados por meio da aplicação de técnicas de aprendizado profundo. Essas técnicas são capazes de considerar e processar volumes de informação superiores à capacidade humana (PICCIALLI et al., 2021).

Neste capítulo, apresentam-se os conceitos necessários ao desenvolvimento deste trabalho: ceratite infecciosa (CI), redes neurais convolucionais (CNN), técnicas de fusão de características, camadas de atenção e o algoritmo Grad-CAM. Em seguida, procede-se à análise dos modelos empregados para a classificação de CI em imagens microscópicas confocais *in vivo*.

2.1 Ceratite infecciosa

A ceratite infecciosa (CI) é caracterizada por uma inflamação da córnea, causada por agentes biológicos microscópicos, como bactérias, vírus, fungos e protozoários. Esta condição representa uma das principais causas de cegueira, tanto em países desenvolvidos quanto em nações emergentes, devido à sua alta morbidade ocular (VERMA et al., 2021).

Como apresentado na introdução deste trabalho, a incidência anual da CI varia entre 2,5 e 799 casos por 100.000 habitantes, dependendo da localização geográfica e dos estudos epidemiológicos (VERMA et al., 2021). O risco é especialmente elevado em regiões com recursos sanitários limitados, onde traumas oculares associados a atividades agrícolas e o acesso reduzido a cuidados médicos são mais frequentes.

O presente estudo concentra-se em dois dos tipos mais comuns de ceratite infecciosa: a ceratite fúngica e a ceratite causada por protozoários do gênero *Acanthamoeba* spp.

A ceratite fúngica representa aproximadamente 6% a 30% dos casos, sendo mais prevalente em áreas tropicais e subtropicais (VERMA et al., 2021). Os principais agentes etiológicos incluem fungos dos gêneros *Aspergillus* e *Fusarium*. Segundo Cabrera-Aguas, Khoo e Watson (2022), os principais fatores de risco envolvem o uso de lentes de contato contaminadas, lesões na córnea, administração de corticosteroides (VERMA et al., 2021) e

condições oculares pré-existentes, como olho seco, blefarite, ceratopatia bolhosa e síndrome de Stevens-Johnson.

Por sua vez, a ceratite protozoária é causada pela contaminação com *Acanthamoeba* spp., geralmente associada ao uso de lentes de contato contaminadas ou ao contato com solo e água contaminados (CABRERA-AGUAS; KHOO; WATSON, 2022). Embora menos prevalente, representando cerca de 5% dos casos, esse tipo de ceratite caracteriza-se por um curso clínico prolongado e de difícil manejo terapêutico, devido à capacidade do protozoário de formar cistos resistentes a agentes antimicrobianos (VERMA et al., 2021).

O diagnóstico da CI pode ser realizado por meio de diferentes técnicas laboratoriais, incluindo colorações específicas, adição de hidróxido de potássio (KOH a 10%), raspagem corneana seguida de cultura microbiológica, reação em cadeia da polimerase (PCR) e microscopia confocal *in vivo* (IVCM). Entre essas, a cultura do raspado da córnea é considerada o método mais assertivo, apesar de seu caráter invasivo e do tempo de espera de aproximadamente 72 horas (SILVA et al., 2025). Como alternativa, as imagens obtidas por IVCM são, posteriormente, analisadas por um profissional da área da saúde e um laudo é emitido confirmando ou não a ceratite infecciosa e seu tipo. Segundo Kryszan et al. (2024), o método não invasivo baseado em imagens de IVCM oferece uma abordagem diagnóstica rápida e com boa acurácia.

Diante da gravidade e da elevada incidência da CI, a inteligência artificial (IA) tem emergido como uma ferramenta promissora na área da saúde, oferecendo novas possibilidades para aprimorar a acurácia e a agilidade dos diagnósticos médicos (VAZQUEZ, 2024). Modelos baseados em IA têm demonstrado desempenho relevante em tarefas relacionadas à análise de imagens médicas, permitindo a extração de informações substanciais por meio de técnicas de visão computacional (KRYSZAN et al., 2024), o que auxilia os profissionais na tomada de decisão clínica.

A base utilizada neste trabalho é a IVCM-Keratitis, disponibilizada publicamente por Essalat et al. (2023). As classes-alvo apresentam assinaturas morfológicas bem definidas em IVCM, como hifas e ramificações fúngicas, bem como estruturas radiais ou dendríticas associadas a amebas, o que motiva a extração de características texturais e de forma em múltiplas escalas. Além disso, a base inclui uma classe em que o tipo de ceratite não é especificado, de modo que essas imagens possam ser encaminhadas à avaliação de um especialista. A metodologia proposta explora representações profundas para capturar tais indícios morfológicos e reduzir a ambiguidade diagnóstica.

2.2 Redes Neurais Convolucionais (CNN)

Na classificação de imagens, métodos estatísticos tradicionais e técnicas de aprendizado de máquina (ML) geralmente dependem de etapas complexas de pré-processamento,

como filtragem, seleção de parâmetros e redução de dimensionalidade (ERSAVAS; SMITH; MATTICK, 2024). Essas etapas têm como objetivo eliminar pequenas variações nos dados, frequentemente associadas a ruídos (artefatos), mas que podem, por vezes, conter informações relevantes para a diferenciação entre classes.

LeCun et al. (1998) afirmam que as redes neurais convolucionais (CNN) são exemplos de arquiteturas que incorporam conhecimento sobre as invariâncias das formas bidimensionais (2D), utilizando padrões locais de conexão e impondo restrições sobre os pesos. Essa estruturação favorece a detecção de padrões espaciais, como bordas, texturas e estruturas repetitivas, diretamente a partir dos dados de entrada.

O surgimento do *deep learning* (DL) representou uma mudança de paradigma na área. Segundo Ersavas, Smith e Mattick (2024), algoritmos baseados em DL, fundamentados em mecanismos de aprendizagem adaptativa, são capazes de capturar tanto padrões lineares quanto não lineares presentes nos dados. Além disso, modelos de DL, como as redes neurais convolucionais (CNNs), têm demonstrado desempenho eficaz na predição de sistemas com comportamento complexo ou caótico (ERSAVAS; SMITH; MATTICK, 2024).

De forma geral, uma CNN moderna é organizada em blocos sucessivos de processamento, que podem incluir: camadas convolucionais, funções de ativação não lineares (como ReLU), camadas de normalização (por exemplo, *batch normalization*), operações de *pooling* ou redução espacial, camadas de regularização (como *dropout*) e, ao final, um cabeçalho de classificação composto por camadas totalmente conectadas e uma função de saída (por exemplo, *softmax*). As primeiras camadas aprendem características de baixo nível (bordas, contrastes locais), enquanto as camadas mais profundas capturam padrões de alto nível, mais relacionados às classes de interesse.

Uma camada convolucional é formada, em essência, por três componentes: os dados de entrada, um conjunto de filtros (ou *kernels*) e os mapas de características (*feature maps*) resultantes. A entrada pode ser uma imagem em escala de cinza ou em múltiplos canais (como RGB ou CMYK). Cada filtro percorre a imagem aplicando uma operação de convolução, com o objetivo de extrair padrões locais relevantes para a tarefa de classificação.

Matematicamente, a saída de uma operação de convolução discreta bidimensional pode ser descrita pela Equação 2.1:

$$S(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} K(m,n) \cdot I(i+m, j+n) \tag{2.1}$$

Na Equação 2.1, os termos são definidos da seguinte forma:

- $S(i,j)$: valor do pixel de saída na posição (i,j) do mapa de características (*feature map*);

- $I(i+m, j+n)$: valor do pixel da imagem de entrada na posição $(i+m, j+n)$;
- $K(m, n)$: valor do peso na posição (m, n) do filtro (*kernel*) de dimensão $M \times N$;
- M e N: dimensões do filtro, representando, respectivamente, a altura e a largura.

Na prática, o filtro é aplicado sobre janelas locais da imagem, deslocando-se horizontal e verticalmente conforme um parâmetro de passo (*stride*) e, opcionalmente, com preenchimento (*padding*) das bordas. Esse processo gera mapas de características em que cada canal representa a resposta de um filtro específico, frequentemente com dimensões espaciais reduzidas em relação à entrada original.

$$y_{i,j} = \max_{(m,n) \in R_{i,j}} x_{m,n} \tag{2.2a}$$

$$y_{i,j} = \frac{1}{|R_{i,j}|} \sum_{(m,n) \in R_{i,j}} x_{m,n} \tag{2.2b}$$

O *pooling* é uma operação responsável por reduzir progressivamente a dimensionalidade espacial das representações intermediárias, preservando as informações mais relevantes e reduzindo a redundância. Ao aplicar funções de agregação, como *max pooling* ou *average pooling*, a rede extrai características mais robustas a pequenas translações, rotações e distorções, favorecendo a generalização do modelo e mitigando o risco de sobreajuste (*overfitting*). O *max pooling* (Equação 2.2a) seleciona o valor máximo de uma região receptiva, enfatizando as características mais salientes, enquanto o *average pooling* (Equação 2.2b) calcula a média, preservando uma representação mais suave e global da informação.

O mapa de características (*feature map*) corresponde, portanto, a uma representação intermediária da imagem, com realce das características locais extraídas pelos filtros. Suas dimensões resultam da combinação dos parâmetros de convolução (tamanho do filtro, *stride*, *padding*) e das operações de *pooling*. Em muitas arquiteturas recentes, em vez de simplesmente aplicar um *flattening* de todos os mapas, utiliza-se uma camada de *global average pooling*, que reduz cada mapa a um único valor médio, produzindo um vetor compacto que alimenta o cabeçalho de classificação. Esse vetor latente é então processado por camadas totalmente conectadas, possivelmente acompanhadas de normalização e *dropout*, até a camada de saída responsável pela predição das classes.

Após a fase de treinamento, a rede neural dispõe de um conjunto de pesos otimizados com base nos vetores de características extraídos das amostras de entrada. Esses pesos permanecem inalterados durante a fase de inferência. Nessa etapa, a rede processa dados inéditos com o objetivo de gerar previsões, aplicando exclusivamente o conhecimento consolidado durante o treinamento, sem qualquer modificação nos parâmetros internos.

No contexto de *transfer learning*, o conhecimento previamente adquirido por um modelo ao ser treinado em um conjunto de dados extenso e de domínio semelhante é aproveitado para resolver um novo problema, possivelmente com um conjunto de dados menor ou mais específico. Essa reutilização de parâmetros permite reduzir o custo computacional, acelerar a convergência do processo de treinamento e alcançar melhor desempenho, especialmente em cenários com limitação de dados. A técnica explora o fato de que as camadas iniciais das redes convolucionais aprendem características genéricas, como bordas e texturas, que podem ser úteis em múltiplos domínios, enquanto as camadas mais profundas capturam padrões mais específicos.

O *fine-tuning* constitui uma extensão do *transfer learning*, na qual apenas parte das camadas da rede é treinável para permitir o reajuste dos pesos durante o novo treinamento. As camadas iniciais permanecem inalteradas para preservar a extração de características genéricas, enquanto as camadas intermediárias ou finais são otimizadas novamente para especializar o modelo às características estatísticas do novo conjunto de dados.

Para a classificação de imagens IVCM, os filtros iniciais capturam microtexturas e bordas das estruturas corneais, ao passo que camadas profundas integram detalhes mais complexos para distinguir as classes da base de dados, conforme discutido no Capítulo **??**. O *transfer learning* acelera a convergência em bases menores, enquanto o *fine-tuning* especializa o modelo à morfologia associada à ceratite infecciosa.

2.3 Modelos de Deep Learning

O ImageNet é um repositório em larga escala voltado para tarefas de visão computacional, construído com base na estrutura semântica do WordNet, um banco lexical que agrupa palavras em conjuntos de sinônimos, denominados *synsets* (DENG et al., 2009). Seu objetivo é associar a cada termo do WordNet entre 500 e 1000 imagens de alta resolução, resultando em milhões de amostras anotadas e organizadas hierarquicamente. Segundo Deng et al. (2009), o ImageNet supera conjuntos de dados anteriores tanto em volume quanto em diversidade, além de apresentar um elevado grau de precisão nas anotações.

Modelos de CNN treinados com o ImageNet têm sido amplamente adotados como ponto de partida para tarefas de classificação, utilizando estratégias de *transfer learning* e *fine-tuning*. Esse aproveitamento de pesos previamente treinados permite acelerar o processo de aprendizagem e adaptar o modelo a domínios específicos, mesmo em conjuntos de dados reduzidos.

Arquiteturas clássicas como AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2017), VGG16 e VGG19 (SIMONYAN; ZISSERMAN, 2014), e ResNet (HE et al., 2016) tornaram-se referências em problemas de classificação, apresentando camadas convolucionais hierárquicas, operações de *pooling* e camadas densamente conectadas. Posteriormente,

redes mais leves e eficientes, como a MobileNet (HOWARD et al., 2017) e a EfficientNet (TAN; LE, 2019), foram propostas visando reduzir o custo computacional, um aspecto essencial em aplicações clínicas em tempo real ou em dispositivos embarcados.

2.3.1 VGG16

A VGG16 é uma rede proposta por Simonyan e Zisserman (2014). A principal característica da VGG16 é a utilização de múltiplas camadas convolucionais com filtros pequenos, de tamanho 3×3, aplicados de forma sequencial (SIMONYAN; ZISSERMAN, 2014). Essa escolha permitiu aumentar a profundidade da rede, mantendo a quantidade de parâmetros e proporcionando uma melhor capacidade de extração de características espaciais. Após cada bloco de camadas convolucionais, a arquitetura inclui camadas de *pooling*, geralmente com janelas de 2×2 e *stride* de 2, com o objetivo de reduzir progressivamente a dimensionalidade dos mapas de características.

Figura 3 – Estrutura da rede neural convolucional VGG16.

A estrutura da VGG16 é composta por cinco blocos convolucionais (Figura 3), seguidos por três camadas totalmente conectadas (*fully connected layers*). Os dois primeiros blocos possuem duas camadas convolucionais cada, enquanto os três blocos subsequentes contêm três camadas convolucionais cada. Todas as camadas convolucionais utilizam a função de ativação ReLU (*Rectified Linear Unit*, Equação 2.3), visando introduzir não-linearidade ao modelo.

$$f(x) = \begin{cases} x, & \text{se } x > 0 \\ 0, & \text{caso contrário} \end{cases} \tag{2.3}$$

A função de ativação ReLU (*Rectified Linear Unit*) aplica uma transformação não linear que anula todas as entradas negativas e preserva os valores positivos.

A saída da última camada convolucional passa por um processo de achatamento (*flattening*), que transforma os mapas de características bi-dimensionais em um vetor unidimensional, no qual é processado pelas camadas totalmente conectadas. A última camada da VGG16 utiliza uma função de ativação *softmax* para produzir uma distribuição de probabilidade entre as classes de saída.

2.3.2 MobileNet-V3

A MobileNet é uma arquitetura de rede neural convolucional desenvolvida especificamente para aplicações de visão computacional em dispositivos móveis e embarcados, onde há severas restrições de memória e capacidade computacional. Desenvolvida inicialmente por (HOWARD et al., 2017), a arquitetura busca um equilíbrio entre acurácia, tamanho do modelo e velocidade de inferência.

Após as versões MobileNetV1 e MobileNetV2, melhorias significativas foram implementadas, resultando na MobileNetV3 (HOWARD et al., 2019), desenvolvida com o auxílio da técnica de busca automática de arquitetura de redes, conhecida como NAS (*Neural Architecture Search*). A abordagem utilizada, denominada *NetAdapt*, permitiu identificar tamanhos de kernel ideais e otimizar a arquitetura para atender a plataformas com recursos computacionais limitados, com foco em desempenho, latência e tamanho do modelo. A rede MobileNetV3 foi projetada para oferecer alta eficiência energética e baixa latência, tornando-a ideal para aplicações em ambientes embarcados. Sua arquitetura está ilustrada na Figura 4.

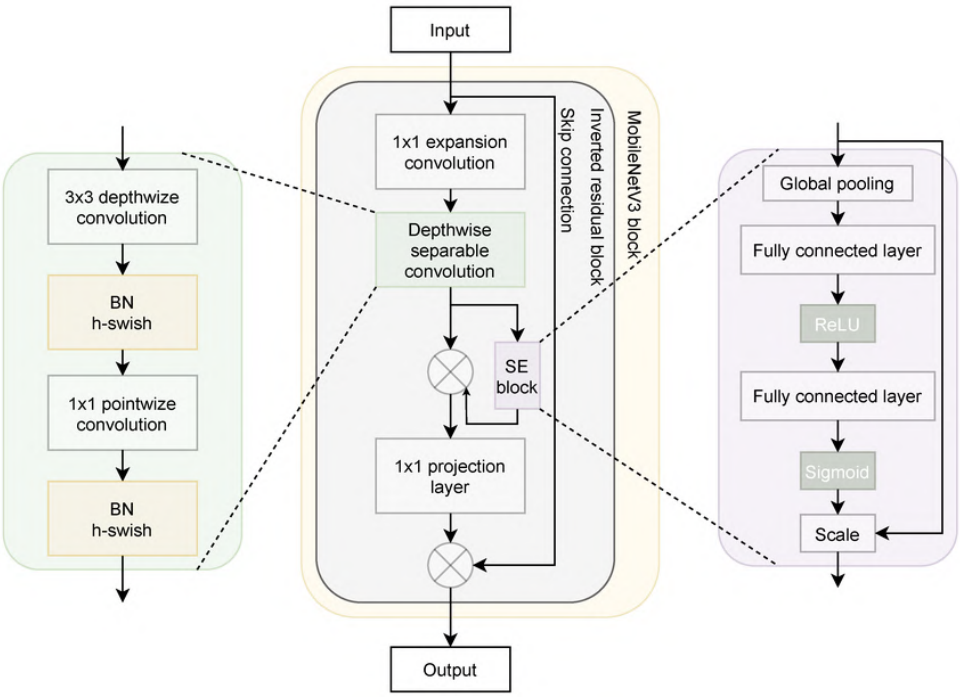


Figura 4 – Arquitetura simplificada do modelo MobileNetV3 (HOWARD et al., 2019).

Conforme ilustrado na Figura 4, a arquitetura da MobileNetV3 é composta por uma estrutura central denominada bloco residual invertido (*inverted residual block*), que integra a operação de convolução separável em profundidade (*depthwise separable convolution*) e o bloco de compressão e excitação (*squeeze-and-excitation* - SE) (HOWARD et al., 2019; TAN et al., 2019).

O bloco residual invertido é inspirado nos blocos gargalo (*bottleneck blocks*), utili-

zando uma conexão residual invertida que liga diretamente as representações de entrada e saída nos mesmos canais. Essa estrutura visa otimizar a extração de características com menor uso de memória.

A convolução separável em profundidade consiste na aplicação de um filtro *depthwise* em cada canal individualmente, seguida por uma convolução *pointwise* (1×1) (HOWARD et al., 2019). Ambas as etapas são acompanhadas de normalização em lote (*batch normalization*) e funções de ativação, como ReLU ou *h-swish*. A operação de soma $\otimes$ consiste na adição ponto a ponto entre dois mapas de características de mesma dimensão.

Essa abordagem substitui a convolução tradicional com o objetivo de reduzir o custo computacional do modelo. Além disso, o bloco treinável *squeeze-and-excitation* (SE) é incorporado para permitir que a rede aprenda a realçar as características mais relevantes de cada canal, ajustando dinamicamente a importância de cada um deles durante o treinamento (HOWARD et al., 2019).

2.3.3 DenseNet-161

A *Dense Convolutional Network* (DenseNet) (HUANG et al., 2017) é uma arquitetura convolucional que se destaca por introduzir conexões diretas entre todas as camadas de um mesmo bloco denso. Diferentemente das redes tradicionais, onde cada camada recebe como entrada apenas a saída da camada anterior, a DenseNet conecta cada camada a todas as anteriores por meio da concatenação de mapas de características (Figura 5). Essa densa conectividade favorece o reaproveitamento de informações e melhora substancialmente o fluxo de gradientes durante o treinamento.

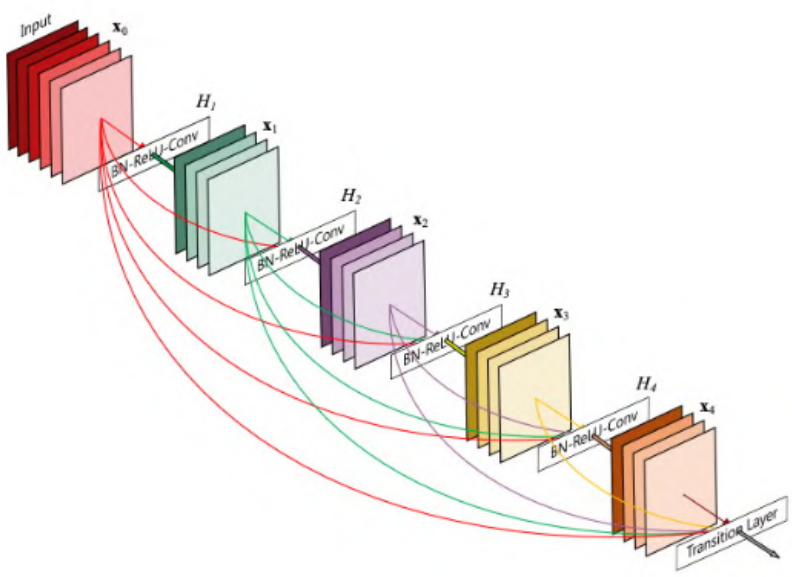


Figura 5 – Estrutura geral da DenseNet, com conexões densas entre as camadas de um mesmo bloco convolucional (HUANG et al., 2017).

Entre suas principais vantagens, destacam-se a eficiência no uso de parâmetros, a mitigação do problema do gradiente que desaparece, o incentivo à reutilização de características e a possibilidade de treinar redes muito profundas com menor risco de sobreajuste (overfitting). Além disso, a DenseNet alcança alto desempenho mesmo com menos parâmetros em comparação com outras arquiteturas profundas, como a ResNet.

Neste trabalho, foi utilizada a variante DenseNet161, selecionada por sua alta capacidade de extração de padrões detalhados em imagens complexas, como as obtidas por microscopia confocal. A escolha também se justifica pela sua ampla adoção na literatura em tarefas de classificação médica, especialmente em estudos relacionados à ceratite infecciosa.

2.3.4 ResNet-101

A ResNet (HE et al., 2015) destacou-se na área de aprendizado profundo ao permitir o treinamento eficiente de redes muito profundas por meio do uso de conexões residuais. Essas conexões, também conhecidas como *skip connections*, consistem em atalhos que somam a entrada de uma camada à sua saída, permitindo que os gradientes fluam diretamente pelas camadas durante o processo de retropropagação (Figura 6).

A estrutura residual resolve o problema da degradação em redes profundas, em que adicionar mais camadas leva a um aumento no erro de treinamento. Com essas conexões, a rede aprende a refinar identidades residuais em vez de mapear diretamente a entrada para a saída, facilitando o aprendizado de funções mais complexas.

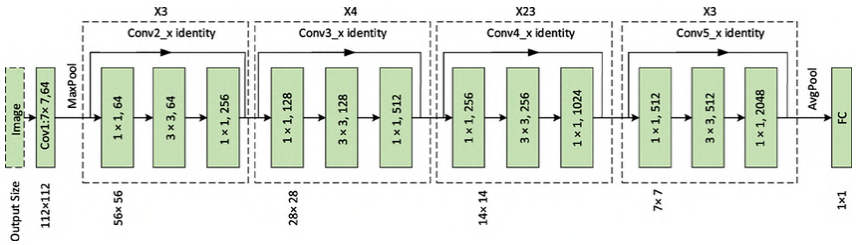


Figura 6 – Arquitetura básica da ResNet, com destaque para os blocos residuais (HE et al., 2015).

Neste trabalho, foi utilizada a versão ResNet101, uma variante com 101 camadas profundas, escolhida por sua eficácia na extração de características profundas e na robustez contra ruídos em imagens médicas. A ResNet101 tem sido amplamente empregada na literatura de classificação de ceratite infecciosa, com resultados expressivos, mesmo em bases de dados reduzidas. Sua profundidade permite capturar representações hierárquicas complexas, sendo especialmente adequada para imagens de microscopia confocal *in vivo* (IVCM), que apresentam estruturas sutis e sobrepostas.

Tais arquiteturas têm se mostrado eficazes na análise de imagens biomédicas, incluindo imagens IVCM voltadas para o diagnóstico de ceratite infecciosa. No entanto,

apesar da robustez dessas redes na extração de padrões locais, as CNNs apresentam limitações relevantes quando se trata de capturar dependências espaciais de longo alcance.

2.4 Self-Attention

Embora eficazes na extração de características locais, as CNNs operam com filtros aplicados a regiões restritas da imagem, o que limita a capacidade de relacionar diretamente regiões distantes. A integração de informações mais amplas ocorre apenas após o empilhamento de múltiplas camadas, o que nem sempre garante uma representação global adequada. Em imagens de IVCM, onde estruturas diagnósticas, como ramificações de fungos ou padrões dendríticos causados por protozoários, podem surgir de forma esparsa e distribuída, essa limitação pode comprometer a identificação adequada das manifestações da ceratite infecciosa.

Para mitigar essas limitações, surgiram abordagens baseadas em mecanismos de atenção, inicialmente desenvolvidas no contexto do processamento de linguagem natural, como nos Transformers (VASWANI et al., 2017). Dentre esses mecanismos, destaca-se o *self-attention* (SA), que permite ao modelo atribuir pesos contextuais a diferentes regiões da entrada, promovendo uma integração explícita entre todas as posições da imagem. Essa capacidade de estabelecer relações globais levou à adaptação dos *Transformers* para o domínio visual, resultando em arquiteturas como os Vision Transformers (ViTs).

Ao permitir que o modelo aprenda a focar seletivamente nas regiões mais relevantes para a tarefa, esses blocos ampliam a sensibilidade a padrões diagnósticos sutis, o que é particularmente desejável em imagens IVCM, caracterizadas por baixo contraste e alta variabilidade morfológica.

Neste trabalho, propõe-se a fusão de vetores de características extraídos de duas arquiteturas convolucionais distintas, VGG16 e MobileNetV3, bem como DenseNet e ResNet, seguida pela aplicação de um bloco de atenção SA. A concatenação dos vetores permite combinar diferentes escalas e profundidades de abstração, capturando aspectos complementares das estruturas corneanas. O mecanismo de atenção atua como um refinador dessa representação conjunta, enfatizando as regiões mais informativas e suprimindo ruídos. Essa estratégia é especialmente útil para lidar com as diferentes manifestações morfológicas da ceratite infecciosa em imagens de IVCM.

Do ponto de vista matemático, seja $\mathbf{X} \in \mathbb{R}^{N \times d}$ o tensor de características resultante da fusão dos vetores extraídos pelas CNNs, onde N representa o número de posições espaciais e d a dimensão do espaço de características. O bloco de *self-attention* projeta $\mathbf{X}$ em três espaços latentes: consultas ($\mathbf{Q}$), chaves ($\mathbf{K}$) e valores ($\mathbf{V}$), por meio de transformações lineares aprendíveis:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V, \tag{2.4}$$

em que $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d_k}$ são matrizes de pesos treináveis e d_k é a dimensão do subespaço latente.

A matriz de pesos de atenção $\mathbf{A} \in \mathbb{R}^{N \times N}$ é então obtida pelo produto escalar entre consultas e chaves, normalizado pela raiz quadrada de d_k e seguido da função *softmax*, de modo a garantir que cada linha de $\mathbf{A}$ represente uma distribuição de probabilidade:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right). \tag{2.5}$$

Por fim, o vetor de saída do bloco de atenção, $\mathbf{Z} \in \mathbb{R}^{N \times d_k}$, é obtido pela combinação linear dos valores ponderados por $\mathbf{A}$:

$$\mathbf{Z} = \mathbf{A}\mathbf{V}. \tag{2.6}$$

Assim, o operador de *self-attention* pode ser sintetizado como:

$$\text{SA}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_Q(\mathbf{X}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{X}\mathbf{W}_V. \tag{2.7}$$

Nesse contexto, cada posição da representação fundida passa a ser atualizada em função de todas as demais, com pesos adaptativos que refletem sua relevância contextual. Essa estratégia é válida para lidar com as diferentes manifestações morfológicas da ceratite infecciosa em imagens de IVCM, nas quais padrões de interesse podem estar distribuídos de forma não contígua.

O próximo capítulo apresenta os trabalhos relacionados, com foco em estudos que empregam imagens de IVCM na inferência automática do agente infeccioso, bem como em técnicas de fusão aplicadas a esse contexto.

3 Trabalhos relacionados

Este capítulo apresenta o estado da arte sobre a aplicação de redes neurais na classificação de imagens obtidas por IVCM. São discutidas abordagens voltadas à classificação automatizada de doenças oculares a partir da análise dessas imagens.

Destaca-se a revisão sistemática conduzida por Kryszan et al. (2024), que analisa propostas baseadas em modelos de inteligência artificial aplicados à microscopia confocal. O estudo enfatiza o uso de CNNs como ferramenta fundamental para aprimorar a acurácia de classificação, especialmente na detecção de patógenos por meio da análise da morfologia das fibras nervosas da córnea. Com base nessa revisão, seis trabalhos se destacam na Tabela 1: Liu et al. (2020), Lv et al. (2020), Xu et al. (2021), Lincke et al. (2023), Liang et al. (2023) e Essalat et al. (2023);

Tabela 1 – Trabalhos relacionados à classificação automática de ceratite utilizando imagens de microscopia confocal.

Autores	**Ano**	**Base de imagens**	**Modelo Proposto**	**Resultados**	**Técnicas adicionais**
Essalat et al. (2023)	2023	4.001	DenseNet161	Acurácia 93,55% Precisão 92,52% Recall 94,77% F1 96,93%	Grad-CAM; *Transfer learning.*
Lincke et al. (2023)	2023	68.970	ResNet101	Acurácia 94,6%	*Transfer learning.*
Liang et al. (2023)	2023	7.278	Híbrido GoogLeNet–VGGNet	Acurácia 97,73% Precisão 98,68% Especificidade 98,54% Recall 97,02% F1 97,84%	Fusão de características.
Lv et al. (2020)	2020	2.088	ResNet101	Acurácia 96,26% Especificidade 98,34% Recall 91,86% AUC 98,75%	*Transfer learning.*
Liu et al. (2020)	2020	1.213	AlexNet	Acurácia 99,95% Recall 99,90% Especificidade 100%	Realce de contraste localizado, fusão por correspondência de histogramas.
Xu et al. (2021)	2021	3.453	ResNet101	Acurácia 96,5% Recall 93,6% Especificidade 98,2% AUC 98,3%	Grad-CAM, Guided Grad-CAM.

O trabalho de Liu et al. (2020) propõe uma estrutura de rede neural convolucional (CNN) utilizando imagens confocais. O método utiliza técnicas de aumento de dados (espelhamento) para balancear um conjunto de dados limitado, seguido de um algoritmo

de pré-processamento chamado *subárea contrast stretching* (SCS), que destaca estruturas-chave, como hifas e esporos. Em seguida, é aplicada a fusão de imagens com base na correspondência de histograma (HMF), combinando as imagens originais com as pré-processadas para realçar detalhes relevantes sem perder a informação original. As imagens fundidas são utilizadas para treinar modelos de CNN com as arquiteturas AlexNet e VGG16. Os resultados experimentais mostram que o modelo AlexNet com fusão HMF obteve uma acurácia de 99,95%, sensibilidade de 99,9% e especificidade de 100%, superando abordagens tradicionais e versões sem fusão. A abordagem destacou-se por aliar desempenho elevado com baixo custo computacional, evidenciando seu potencial para aplicações em tempo real no auxílio ao diagnóstico clínico de FK.

O estudo de Lv et al. (2020) teve como objetivo desenvolver um algoritmo baseado em *deep learning* para a classificação de ceratite fúngica (FK) em imagens de microscopia confocal *in vivo* (IVCM). Foi utilizada a arquitetura ResNet-101, uma rede residual profunda que facilita o treinamento de modelos com muitas camadas por meio de conexões de atalho. Segundo os autores, o sistema proposto não apenas auxilia os médicos no diagnóstico de FK, mas também permite o monitoramento da presença de estruturas fúngicas residuais ao longo do acompanhamento clínico. A validação externa foi conduzida com dados independentes, resultando em uma área sob a curva ROC (AUC) de 98,75%, uma acurácia de 96,26% na detecção de hifas, especificidade de 98,34% e sensibilidade de 91,86%. Além disso, ao incluir pacientes diabéticos de uma base de imagens privada, observou-se a manutenção do desempenho do modelo, com acurácia, sensibilidade e especificidade de 93,64%, 82,56% e 98,89%, respectivamente.

Xu et al. (2021) avaliou a utilidade clínica de um sistema de inteligência artificial explicável (XAI) para a classificação de ceratite fúngica (FK) com base em imagens de IVCM. Os autores treinaram um modelo de *deep learning* com a arquitetura ResNet101 em 2.088 imagens rotuladas e o testaram em um conjunto de 1.089 imagens. O modelo atingiu acurácia de 96,5%, sensibilidade de 93,6%, especificidade de 98,2% e AUC de 0,983. Outra contribuição relevante foi a integração da explicabilidade por meio dos métodos *Grad-CAM* e *Guided Grad-CAM*, permitindo visualizar as regiões da imagem mais relevantes para a predição do modelo. Para consolidar os resultados, foi realizado um ensaio clínico com nove oftalmologistas (experientes, competentes e novatos) para análise das imagens sob três condições: sem assistência, com assistência de IA e com XAI. A XAI aumentou significativamente a acurácia, especialmente para leitores menos experientes (de 80,7% para 90,7%), sem aumento no tempo de leitura. Os autores concluem que a explicabilidade não só melhora o desempenho do diagnóstico, como também oferece uma ferramenta de apoio pedagógico e clínico, promovendo confiança no uso de sistemas automatizados em contextos reais de decisão médica.

O estudo conduzido por Lincke et al. (2023) teve como objetivo desenvolver um

sistema de apoio à decisão baseado em agentes de IA para a classificação de ceratite por *Acanthamoeba* (AK), por meio da implementação de modelos de *deep learning*, técnicas de processamento de imagens e contribuições de especialistas. A validação cruzada ($k = 5$) foi conduzida em um conjunto de 47.734 imagens obtidas em uma clínica oftalmológica na Suécia. O sistema recebe imagens de IVCM como entrada, classifica e remove artefatos com acurácia de 93%, e em seguida categoriza as imagens remanescentes segundo as camadas da córnea, com acurácia de 90%. Posteriormente, o modelo determina se a córnea está saudável ou apresenta sinais de infecção, com acurácia de 95%. Em casos positivos, identifica imagens com indícios de ceratite por *Acanthamoeba* (AK) com acurácia de 85% e localiza a região de interesse com acurácia de 75,5% (LINCKE et al., 2023).

O trabalho de Liang et al. (2023) propõe um método automatizado de classificação para o diagnóstico de ceratite fúngica (FK) utilizando imagens de IVCM. A arquitetura adota uma abordagem de dois fluxos, concatenando os vetores de características das redes GoogLeNet e VGGNet. O fluxo principal processa a imagem original para a extração de características em nível global, enquanto o fluxo auxiliar recebe como entrada uma representação da estrutura aproximada das hifas, obtida pela subtração da média dos valores dos pixels da imagem original, incorporando esse conhecimento prévio ao processo de aprendizagem. O objetivo do fluxo auxiliar é reforçar a discriminação e o realce das regiões contendo hifas, complementando as informações extraídas pelo fluxo principal. As características provenientes de ambos os fluxos são integradas por concatenação na dimensão dos canais e utilizadas para a classificação binária das imagens (normais ou com presença de hifas). Avaliado no conjunto de teste, o modelo alcançou acurácia de 97,73%, precisão de 98,68%, sensibilidade de 97,02%, especificidade de 98,54% e *F1-score* de 97,84%.

No estudo de Essalat et al. (2023), foi introduzida a base de dados IVCM-Keratitis, também utilizada neste trabalho, composta por 4001 imagens obtidas por microscopia confocal *in vivo* (IVCM), categorizadas em quatro classes: ceratite por *Acanthamoeba* (AK), ceratite fúngica (FK), ceratite não especificada (NSK) e olhos saudáveis. O trabalho avaliou dez arquiteturas de redes neurais convolucionais (CNNs) com o objetivo de aprimorar a acurácia da classificação de ceratite infecciosa. Dentre os modelos testados, o DenseNet161 obteve o melhor desempenho, com acurácia de 93,55%, precisão de 92,52%, *recall* de 94,77% e *F1-score* de 96,93%. Além disso, foram aplicados mapas de calor que destacam regiões relevantes da imagem para a tomada de decisão do modelo (*GradCAM*). O estudo reforça o potencial de modelos baseados em *deep learning* na diferenciação entre os tipos de ceratite infecciosa, evidenciando sua aplicabilidade como ferramenta auxiliar no diagnóstico clínico (ESSALAT et al., 2023).

3.1 Discussão da Revisão de Literatura

Os trabalhos revisados convergem no uso de técnicas de *deep learning* para a classificação de ceratites infecciosas a partir de imagens obtidas por IVCM. No entanto, apresentam distintas abordagens quanto ao escopo clínico, às arquiteturas empregadas, às estratégias de validação e as bases de dados, sendo estas, em sua maioria, privadas e de acesso restrito, o que limita a reprodutibilidade e dificulta comparações diretas entre modelos.

O estudo de Lincke et al. (2023), por exemplo, propõe um pipeline diagnóstico completo para ceratite por Acanthamoeba (AK), com segmentação em camadas da córnea e remoção de artefatos, validado em um banco institucional com 47.734 imagens. Já Liang et al. (2023) e Lv et al. (2020) concentram-se na detecção de ceratite fúngica (FK), utilizando redes convolucionais para capturar estruturas de hifas, mas sem realizar validação cruzada ou testes externos, fator que pode inflar artificialmente as métricas. Destaca-se o trabalho de Liu et al. (2020), o único a explorar uma arquitetura de duplo-fluxo (AlexNet + VGG16), ainda que sem mecanismos de atenção ou de explicabilidade.

Em uma linha complementar, Xu et al. (2021) introduz mapas Grad-CAM e Guided Grad-CAM para explorar a interpretabilidade das decisões do modelo, demonstrando ganhos relevantes para clínicos não especializados. Por fim, Essalat et al. (2023) apresentam a única base de dados pública conhecida até o momento, a IVCM-Keratitis, e testam diferentes modelos, sendo que o DenseNet alcança um desempenho competitivo, com uma taxa F1-score de 93,5%.

Apesar dos avanços, lacunas metodológicas persistem: ausência de validação externa, uso limitado de técnicas de explicabilidade, baixa reprodutibilidade e escassez de abordagens de multi fluxo eficientes. Além disso, muitos estudos relatam acurácias elevadas (próximas a 99%), sem explorar criticamente fatores como *data leakage*, ajuste de hiperparâmetros, particionamentos enviesados ou conjuntos pequenos.

Neste cenário, o presente estudo propõe uma abordagem original e reprodutível, explorando a base pública IVCM-Keratitis, previamente introduzida por Essalat et al. (2023). Esta base foi selecionada por sua disponibilidade e viabiliza avaliações comparáveis entre arquiteturas de *deep learning*. A partir dela, desenvolvemos modelos de híbridos, nos quais se concatenam os vetores de características extraídos pelas redes VGG16, MobileNetV3, DenseNet161 e ResNet101 e, posteriormente, são refinados por um bloco de atenção do tipo *self-attention*, inspirado na arquitetura dos Transformers.

Os resultados foram avaliados por meio de validação cruzada estratificada em cinco *folds*, mantendo a proporção entre as classes. A métrica AUC foi adotada como referência para a sensibilidade global do modelo, além da taxa F1-score, acurácia e precisão, que serão detalhadas no próximo capítulo. Em termos de interpretabilidade, a técnica Grad-CAM é

utilizada para gerar mapas de ativação visual, permitindo identificar as regiões da imagem que mais influenciam a decisão da rede. Essa transparência é especialmente relevante em aplicações clínicas, como evidenciado por Xu et al. (2021), e fortalece a confiabilidade do sistema proposto.

Ainda que o presente trabalho esteja ancorado em arquiteturas convolucionais, reconhece-se a crescente ascensão de modelos baseados em *Transformers*, como o Vision Transformer (ViT) e o Swin Transformer, os quais representam uma linha promissora para investigações futuras, sobretudo pela sua capacidade de capturar relações espaciais globais que as CNNs tradicionais tendem a perder.

4 Materiais e Métodos

Este capítulo descreve o delineamento experimental adotado para o treinamento e a avaliação de modelos de *deep learning* voltados à classificação de imagens IVCM nas classes AK, FK, NSK e Normal. A metodologia foi estruturada para assegurar reprodutibilidade, comparabilidade entre configurações e uma estimativa imparcial de generalização. Para isso, aplicam-se técnicas de transformação às imagens originais com o objetivo de gerar amostras sintéticas e mitigar o desequilíbrio entre classes, enquanto se mantém um conjunto de teste composto exclusivamente por imagens originais, completamente isolado das etapas de ajuste de pesos e seleção de modelo. Sobre o conjunto de treinamento, emprega-se validação cruzada *Stratified K-fold*, garantindo partições representativas e preservando, em cada *fold*, a composição definida entre amostras originais e sintéticas. Adicionalmente, blocos de *Self-Attention* são incorporados às arquiteturas convolucionais, originando variantes dos modelos base, as quais são comparadas por meio de métricas quantitativas de desempenho. Ao longo do capítulo, detalham-se a base IVCM-Keratitis (ESSALAT et al., 2023), os procedimentos de pré-processamento e aumento de dados, o protocolo de validação, as arquiteturas avaliadas (com e sem atenção) e as métricas utilizadas para a avaliação final.

4.1 Conjunto IVCM-Keratitis

A base IVCM-Keratitis é composta por 2.155 imagens em formato JPEG, com resolução de 768×576 pixels, organizadas em quatro classes: 891 imagens de ceratite por *Acanthamoeba* (AK), 600 imagens de ceratite fúngica (FK), 182 imagens de ceratite não especificada (NSK) e 482 imagens de olhos saudáveis.

As imagens foram coletadas entre 2008 e 2018 no Banco Central de Olhos, no Irã, utilizando o equipamento de varredura confocal ConfoScan 3.0, desenvolvido pela empresa italiana Nidek (ESSALAT et al., 2023). Durante o processo de consolidação, a rotulagem das imagens foi confirmada por exames microbiológicos de raspagem e cultura, e as informações clínicas associadas foram compatíveis com os respectivos diagnósticos. Além disso, as imagens foram avaliadas de forma independente por dois médicos-cientistas, sendo excluídas aquelas de baixa qualidade, como imagens desfocadas, excessivamente claras, escuras ou fora de foco.

As imagens que compõem a base de dados apresentam assinaturas morfológicas distintas para cada agente etiológico, conforme ilustrado na Figura 7.

No exemplo de ceratite por *Acanthamoeba* (AK), o diagnóstico por IVCM baseia-se na identificação de estruturas circulares ou ovais de alta refletividade, correspondentes

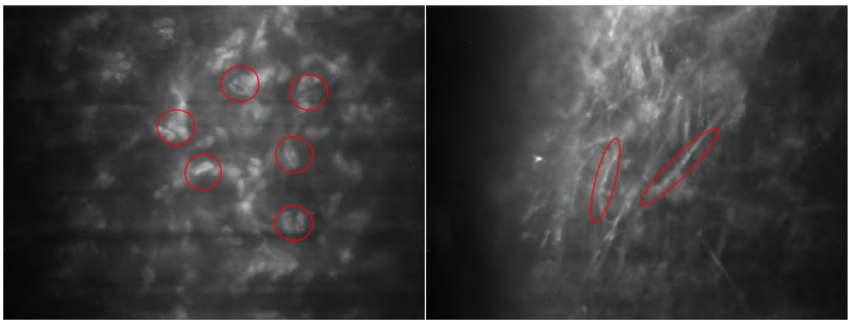


Figura 7 – Exemplos de imagens IVCM evidenciando as diferenças morfológicas entre patógenos: à esquerda, possíveis cistos em um caso de AK; à direita, possíveis hifas em um caso de FK.

aos cistos do protozoário. Já na ceratite fúngica (FK), a caracterização visual é definida pela presença de estruturas filamentosas, lineares e ramificadas, que representam as hifas fúngicas infiltradas no estroma corneal.

4.2 Modelo de Classificação Utilizados

As arquiteturas VGG16 Simonyan e Zisserman (2014), MobileNetV3-Large (HOWARD et al., 2019), DenseNet (HUANG et al., 2017) e ResNet (HE et al., 2015) foram selecionadas como modelos base para a tarefa de classificação de imagens obtidas por microscopia confocal. A VGG16, em sua configuração clássica, foi escolhida pela capacidade de extrair características detalhadas por meio de convoluções sequenciais. A MobileNetV3-Large foi adotada por seu baixo custo computacional e boa relação entre desempenho e eficiência, sendo modificada neste trabalho com a adição de um bloco de *self-attention* ao final do extrator de características, com o objetivo de destacar automaticamente as regiões mais relevantes no vetor de características. As arquiteturas DenseNet e ResNet foram incluídas por sua ampla adoção na literatura voltada à classificação de ceratite infecciosa, além de sua reconhecida capacidade de capturar padrões visuais profundos por meio de conexões densas e residuais, que favorecem o fluxo de gradientes e o reaproveitamento de informações ao longo das camadas.

4.3 Técnicas de Pré-processamento

Nesta seção, são descritas as técnicas de pré-processamento, aplicadas somente às imagens de treinamento, com o intuito de aumentar a quantidade de imagens disponíveis e promover uma maior variabilidade no conjunto. Tais transformações incluem: *color jitter*, desfoque gaussiano (*Gaussian blur*) e espelhamento horizontal e vertical (*flip*). A escolha dessas técnicas foi parcialmente inspirada nas abordagens adotadas por Essalat et al. (2023).

A rotação é uma técnica de aumento de dados amplamente utilizada em imagens médicas, que consiste em aplicar uma transformação angular à imagem original. Ao rotacionar imagens dentro de um intervalo definido em ângulos de -35° a +35°, o modelo torna-se mais robusto a variações de orientação, que são comuns em exames de microscopia ou em cortes histológicos. Essa técnica é especialmente relevante em contextos onde a orientação da amostra pode variar entre aquisições, como em lâminas de tecidos ou registros confocais (COSSIO, 2023).

A técnica de *color jitter* consiste em introduzir variações aleatórias nos atributos de cor da imagem, como brilho, contraste, saturação e matiz. No contexto da patologia computacional e de exames como a microscopia confocal, essas alterações simulam variações nas condições de iluminação, coloração de tecidos e configurações de captura. Essa técnica fortalece a robustez do modelo, permitindo que ele reconheça padrões mesmo quando as características visuais são levemente modificadas. Além disso, a *color jitter* é útil para modelar a variabilidade de imagens presentes nos bancos de dados médicos, promovendo maior interoperabilidade dos sistemas de IA treinados (COSSIO, 2023).

O desfoque gaussiano é uma técnica de filtragem que suaviza a imagem ao aplicar um filtro com distribuição gaussiana sobre os pixels. Essa operação atenua componentes de alta frequência, como ruídos e bordas muito acentuadas, gerando uma imagem mais suave. Em imagens médicas, como aquelas obtidas por microscopia, o desfoque gaussiano pode simular imperfeições de aquisição ou desfoques reais decorrentes do equipamento. Ao treinar modelos com versões suavizadas das imagens, melhora-se a capacidade do sistema de manter o desempenho em imagens com diferentes níveis de ruído, o que é crucial para aplicações robustas em ambientes clínicos (COSSIO, 2023).

O espelhamento, tanto horizontal quanto vertical, é uma técnica eficaz de aumento de dados que consiste em inverter a imagem ao longo do eixo horizontal ou vertical. Essa técnica aumenta a diversidade do conjunto de dados e reduz a sensibilidade do modelo à orientação da amostra. No caso de imagens médicas, como as de tecidos corneanos ou exames confocais, o espelhamento pode ajudar o modelo a reconhecer estruturas simétricas ou padrões que ocorrem em múltiplas direções. Quando ambas as direções são aplicadas em conjunto, há um aumento expressivo na variabilidade dos dados disponíveis para treinamento, sem introdução de artefatos ou distorções estruturais (COSSIO, 2023).

4.4 Validação Cruzada K-Fold

No desenvolvimento de modelos de *deep learning*, o processo de treinamento é uma etapa fundamental. Para que esse processo seja realizado de forma eficaz, é essencial considerar a disposição das amostras e seus respectivos rótulos. Conjuntos de dados desbalanceados, ou seja, com distribuição não uniforme entre as classes, podem induzir

o modelo a um aprendizado enviesado, resultando em *overfitting*. Portanto, é necessário promover o balanceamento dos dados, a fim de garantir que o modelo seja capaz de generalizar adequadamente sobre todas as classes envolvidas.

Uma técnica amplamente utilizada para obter uma estimativa mais robusta do desempenho é a validação cruzada via *K-fold*. Nesse procedimento, o conjunto de dados é particionado em k subconjuntos (*folds*) de tamanhos aproximadamente iguais; a cada iteração, um *fold* é reservado para teste e os demais são usados para treinamento, de modo que todas as amostras participem do processo de avaliação em algum momento (PRUSTY; PATNAIK; DASH, 2022).

Em problemas de classificação com desequilíbrio entre classes, entretanto, o *K-fold* tradicional pode gerar partições pouco representativas (por exemplo, *folds* com escassez ou ausência da classe minoritária), tornando a avaliação instável e potencialmente enviesada. Por essa razão, emprega-se comumente a validação cruzada **estratificada** (*Stratified K-fold*), na qual se busca preservar, em cada *fold*, proporções de classes semelhantes às do conjunto original. Isso não elimina o desbalanceamento em si, mas torna a estimativa de desempenho mais confiável e comparável ao longo das iterações, especialmente quando reportada com métricas sensíveis à classe minoritária (F1-score, *recall* e AUC).

Durante a divisão dos dados, a estratificação assegura a preservação da proporção original das classes em cada subconjunto. Essa abordagem, utilizada neste trabalho, contribui para reduzir o viés durante o treinamento, mitigar o risco de sobreajuste e promover avaliações mais robustas e representativas do desempenho do modelo (PRUSTY; PATNAIK; DASH, 2022).

Neste estudo, a validação cruzada é empregada na etapa de treinamento e seleção de modelo, sem utilizar o conjunto de teste em nenhuma decisão de ajuste. Inicialmente, a base IVCM-Keratitis é particionada de forma estratificada em dois subconjuntos mutuamente exclusivos: (i) um conjunto de treinamento (*treino*) e (ii) um conjunto de teste (*teste*), que permanece totalmente isolado durante o ajuste de pesos.

Em seguida, a validação cruzada *Stratified K-fold* é aplicada **apenas** ao conjunto de *treino*. Especificamente, o *treino* é subdividido em k *folds* estratificados e, em cada iteração, $k-1$ *folds* são utilizados para o treinamento do modelo, enquanto o *fold* restante é utilizado como validação interna para monitoramento do treinamento (por exemplo, *early stopping*). Ao final das k iterações, a configuração final é definida a partir do desempenho médio na validação cruzada e, então, o modelo é re-treinado utilizando todo o conjunto de *treino*. Por fim, o desempenho é reportado por meio de uma **única** avaliação no conjunto de *teste*, garantindo uma estimativa imparcial de generalização.

A base de dados foi previamente balanceada pelas transformações de aumento de dados descritas anteriormente, de forma a tornar aproximadamente equivalente (25%)

a participação de cada classe no conjunto total. A partir dessa base balanceada, cada *fold* foi construído de modo que as imagens sintéticas geradas pelo aumento de dados fossem alocadas exclusivamente ao subconjunto de treinamento, enquanto os conjuntos de validação e teste permaneceram compostos apenas por imagens originais. Dessa forma, preserva-se a distribuição global das classes e, ao mesmo tempo, evita-se que variações artificiais influenciem diretamente a etapa de avaliação, reduzindo o risco de superestimar o desempenho do modelo.

4.5 Métricas para Avaliação

Nesta seção são apresentadas as métricas adotadas para a avaliação quantitativa dos modelos de classificação. A escolha dessas medidas é fundamental para garantir uma interpretação fiel do desempenho, sobretudo em cenários multiclasses e potencialmente desbalanceados, nos quais uma única métrica pode produzir conclusões enviesadas. Assim, foram utilizadas as métricas de acurácia, precisão, sensibilidade (*recall*) e *F1-score*, a fim de caracterizar tanto a taxa global de acertos quanto o comportamento do modelo por classe, evidenciando compromissos entre erros do tipo falso positivo e falso negativo.

A acurácia representa a proporção de predições corretas realizadas pelo modelo em relação ao total de instâncias avaliadas (GRANDINI; BAGLI; VISANI, 2020). Em termos práticos, corresponde à soma dos valores da diagonal principal da matriz de confusão dividida pelo total de amostras. Embora seja uma métrica intuitiva e fácil de calcular, a acurácia pode ser enganosa em cenários com desbalanceamento de classes, pois tende a privilegiar o desempenho em classes majoritárias, ocultando falhas graves nas classes minoritárias (GRANDINI; BAGLI; VISANI, 2020).

A métrica de precisão quantifica a confiabilidade do modelo ao prever uma determinada classe, sendo definida como a razão entre verdadeiros positivos e o total de predições positivas (verdadeiros positivos + falsos positivos). Em outras palavras, mede a proporção de instâncias que foram corretamente classificadas como pertencentes a uma classe, dentre todas as instâncias que o modelo rotulou como pertencentes a ela (GRANDINI; BAGLI; VISANI, 2020).

A sensibilidade (*recall*) mede a capacidade do modelo de identificar corretamente todas as instâncias reais de uma determinada classe (GRANDINI; BAGLI; VISANI, 2020). É calculada como a razão entre verdadeiros positivos e o total de instâncias da classe real (verdadeiros positivos + falsos negativos). Essa métrica é essencial quando o objetivo é capturar todas as ocorrências de um fenômeno, mesmo que isso implique gerar falsos positivos. Em contextos como a detecção de doenças, a sensibilidade é crítica, pois penaliza o modelo por ignorar casos reais (GRANDINI; BAGLI; VISANI, 2020).

O *F1-Score* é a média harmônica entre precisão e sensibilidade, oferecendo um

equilíbrio entre essas duas métricas. O F1-Score favorece modelos que conseguem manter ambos os valores em níveis altos e similares, penalizando fortemente os casos em que uma das métricas é baixa (GRANDINI; BAGLI; VISANI, 2020).

As métricas apresentadas foram empregadas neste trabalho para avaliar quantitativamente o desempenho dos modelos nas fases de treinamento e teste. A partir desses indicadores, foi possível realizar comparações consistentes entre as diferentes abordagens propostas, identificando aquela que obteve o melhor equilíbrio entre as medidas de acurácia, precisão, sensibilidade e *F1-score*.

4.5.1 Análise de Complexidade

Com base na identificação da abordagem com melhor desempenho, realizou-se uma análise comparativa das características computacionais e físicas dos modelos utilizados na literatura. Para comparação, além dos modelos citados na Seção 3, foram incluídos modelos como *Vision Transformer*, VGG19, ConvNeXt e InceptionV3. A avaliação considerou métricas como número de parâmetros, tamanho físico (em *megabytes*) e taxa de operações em ponto flutuante por segundo (FLOPs). Essa análise reforça a capacidade da abordagem proposta em alcançar resultados expressivos, mesmo com menor demanda computacional.

5 Metodologia

Este trabalho propõe a criação de uma arquitetura híbridas baseada na exploração de combinações de redes neurais convolucionais para a tarefa de classificação de imagens IVCM.

São empregadas as arquiteturas convolucionais VGG16, MobileNetV3, DenseNet e ResNet. A VGG16, em sua configuração clássica, foi escolhida pela capacidade de extrair características detalhadas por meio de convoluções sequenciais. A MobileNetV3-Large foi adotada por seu baixo custo computacional e sua boa relação entre desempenho e eficiência, sendo modificada neste trabalho com a adição de um bloco de *self-attention* ao final do extrator de características, com o objetivo de destacar automaticamente as regiões mais relevantes no vetor de características.

As arquiteturas DenseNet e ResNet foram incluídas por sua ampla adoção na literatura voltada à classificação da ceratite infecciosa, além de sua reconhecida capacidade de capturar padrões visuais profundos por meio de conexões densas e residuais, que favorecem o fluxo de gradientes e o reaproveitamento de informações ao longo das camadas.

A metodologia aplicada está representada pela Figura 8 e é dividida em cinco etapas: estruturação do conjunto de dados (I), aumento artificial de dados (II), modelo proposto (III) e resultados das métricas (IV).

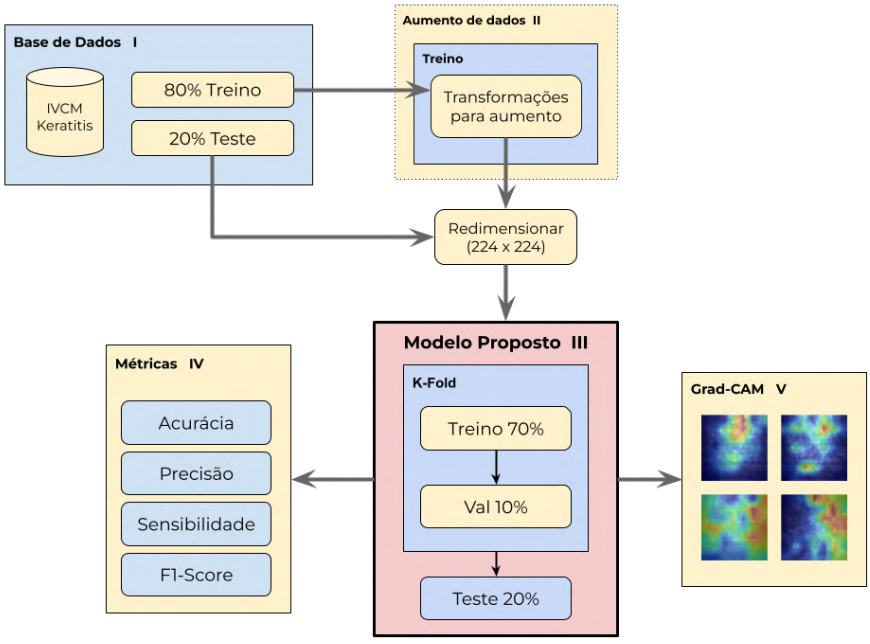


Figura 8 – Metodologia composta pelas etapas apresentadas.

As abordagens implementadas são baseadas em arquiteturas convolucionais previamente definidas, cujas saídas convolucionais finais são convertidas em vetores de características lineares. O bloco de atenção SA é empregado nessas abordagens, visando explorar diferentes formas de refinamento de características.

5.1 Conjunto de dados

As imagens foram disponibilizadas publicamente por Essalat et al. (2023). Inicialmente, o conjunto de dados foi dividido em dois subconjuntos, com 80% das imagens destinadas ao treinamento e 20% ao teste, preservando a proporção entre as classes por meio de uma divisão estratificada. Em seguida, na etapa II, as classes do conjunto de treino ceratite fúngica (FK), ceratite por *Acanthamoeba* (AK), ceratite não especificada (NSK) e córnea normal (Normal) foram balanceadas por meio de técnicas de aumento de dados. Este aumento do conjunto de treinamento está representado pela Figura 9.

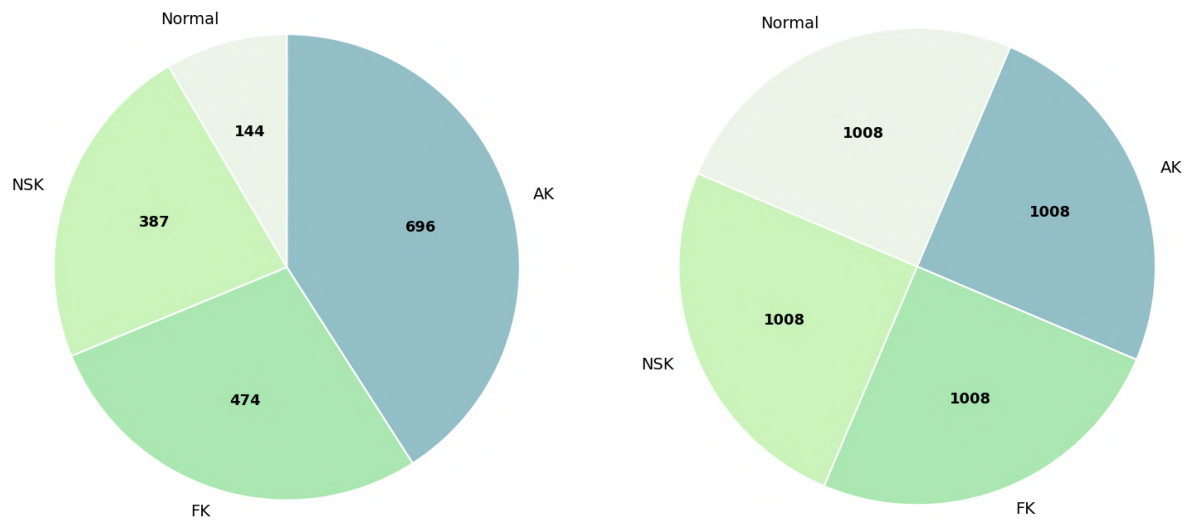
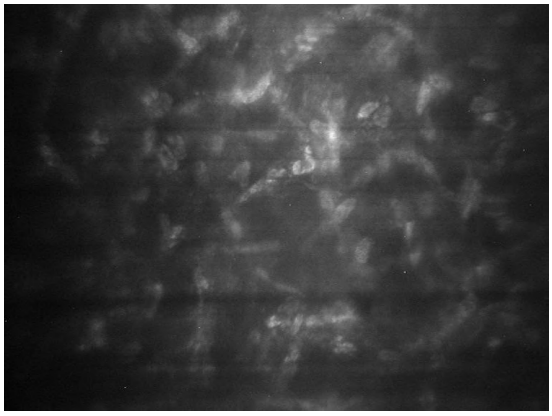
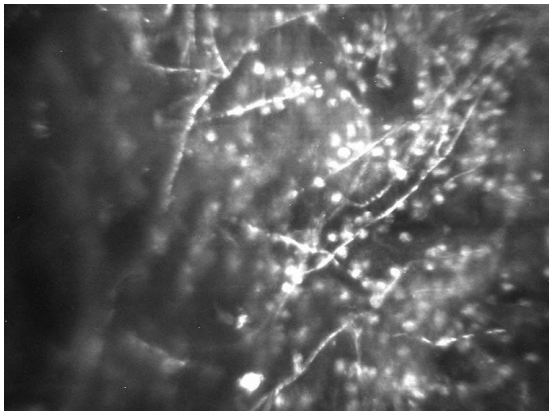


Figura 9 – Distribuição das imagens de treinamento por classe (a) antes e (b) após o balanceamento do conjunto de dados. Após o balanceamento, a base de dados consta com 4032 imagens para treino (88,7%) e 454 (12,3%) imagens de teste, totalizando 4.486 imagens.

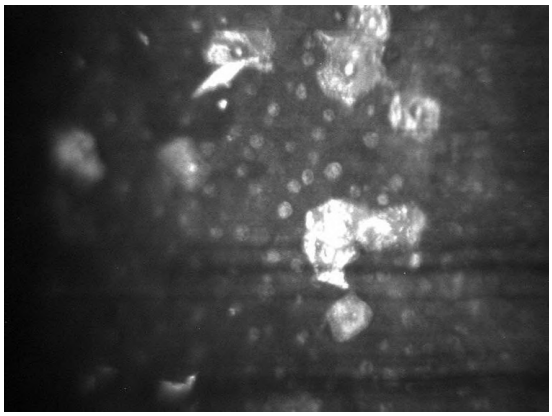
É executado um algoritmo para a detecção e remoção de imagens duplicadas (pixel a pixel). O objetivo é evitar redundâncias e potenciais vieses durante o treinamento do modelo. Além disso, foram desconsideradas imagens com elevado nível de ruído, que poderiam comprometer a qualidade do aprendizado. A Figura 10 representa amostras que foram consideradas válidas e a Figura 11 apresenta exemplos com alto nível de ruído, que foram removidos do conjunto de dados.



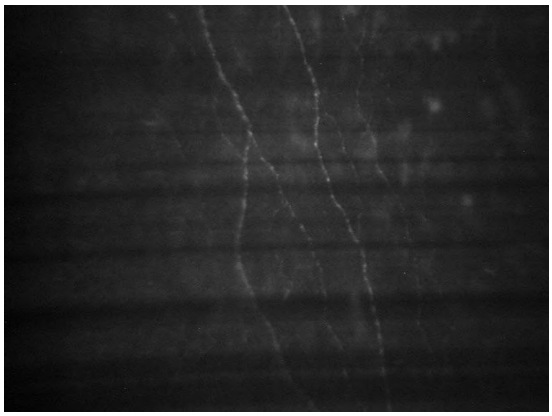
(a) Classe AK



(b) Classe FK



(c) Classe NSK

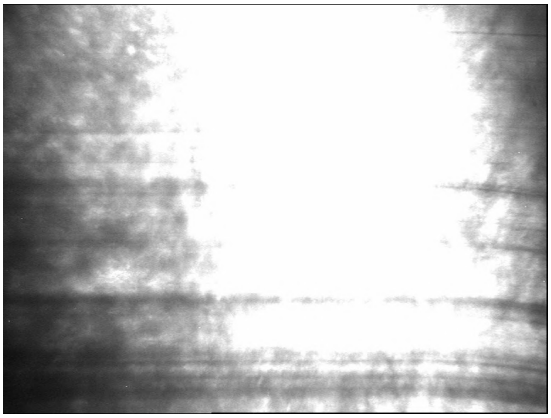


(d) Classe Normal

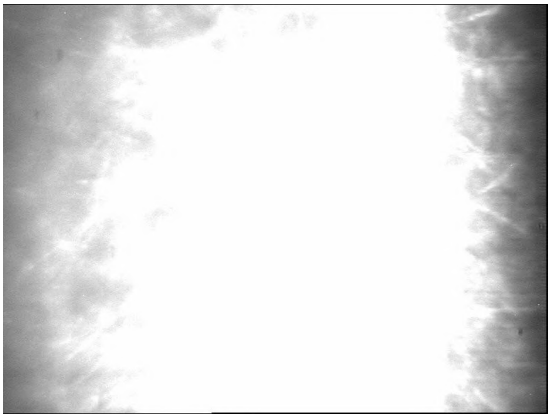
Figura 10 – Exemplos representativos válidos das quatro classes do conjunto de dados utilizado. Individualmente, as características estão sendo representadas de forma clara. Em (a) é possível observar protozoários do gênero *Acanthamoeba*. Em (b) as estruturas de hifas estão sendo destacadas na imagem, característica proveniente dos fungos. Na amostra (c), o autor Essalat et al. (2023) considera um organismo que causa ceratite, porém não especificado (NSK). Em (d) apresenta uma córnea saudável.

Seis técnicas de aumento de dados foram aplicadas exclusivamente ao conjunto de treinamento, com o objetivo de reduzir o desbalanceamento entre as classes AK, FK, NSK e Normal: *horizontal flip*, *vertical flip*, *color jitter*, *Gaussian blur*, rotação no sentido horário e rotação no sentido anti-horário (com ângulos entre 5° e 35°). Para cada imagem original, foram geradas até seis variantes distintas, resultando em um máximo de sete instâncias por amostra (1 original e 6 sintéticas), em um processo de aumento físico (*offline*) com armazenamento local.

Após o aumento, definiu-se como tamanho-alvo para cada classe o valor de 1.008 imagens. Esse valor decorre do fato de que a classe minoritária NSK possui 144 imagens originais e pode gerar, no máximo, $144 \times 6 = 864$ imagens sintéticas por meio das seis transformações consideradas, totalizando $144 + 864 = 1.008$ instâncias. Assim, todas as 144 imagens originais de NSK são mantidas e complementadas por seis versões aumentadas de

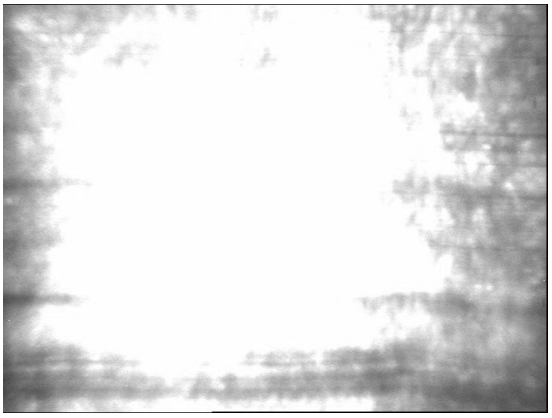


(a) Classe AK



(b) Classe FK

(c) Classe NSK



(d) Classe Normal

Figura 11 – Amostras excluídas do processo de treinamento neste projeto. As imagens das classes representadas em (a), (b) e (d) apresentam um nível excessivo de luminosidade, o que pode ofuscar estruturas relevantes e introduzir viés na etapa de aprendizado do modelo. Já as imagens ilustradas em (c), com baixa luminosidade e contraste, dificultam a visualização de características importantes. Ambas as situações foram identificadas em diferentes classes do conjunto de dados.

cada amostra, atingindo exatamente 1.008 imagens nessa classe.

Para as classes AK, FK e Normal, que inicialmente possuíam mais amostras do que NSK, o procedimento foi análogo: primeiro, todas as imagens originais foram preservadas e submetidas ao mesmo conjunto de transformações; em seguida, as imagens sintéticas excedentes foram removidas por amostragem aleatória até que cada classe atingisse também 1.008 instâncias. Dessa forma, o conjunto de treinamento torna-se balanceado, preservando integralmente os dados reais e limitando o número de cópias aumentadas, o que reduz o risco de redundância excessiva. Por fim, todas as imagens dos conjuntos de treinamento, validação e teste foram redimensionadas para 224×224 pixels.

A estratégia de aumento de dados adotada neste trabalho é inspirada na metodologia proposta por (ESSALAT et al., 2023), mas com modificações relevantes. Diferentemente

do estudo original, as transformações foram executadas de forma pré-processada (*offline*) e as rotações foram aplicadas, separando explicitamente ângulos positivos e negativos.

As Figuras 12 e 13 ilustram amostras sintéticas obtidas após a aplicação das técnicas de aumento de dados. As transformações aplicadas contribuem para a ampliação da variabilidade intra-classe, simulando diferentes condições de captura e características morfológicas. Essa diversidade artificialmente induzida é essencial para promover a robustez do modelo, reduzindo o risco de *overfitting* e melhorando sua capacidade de generalização em cenários clínicos reais.

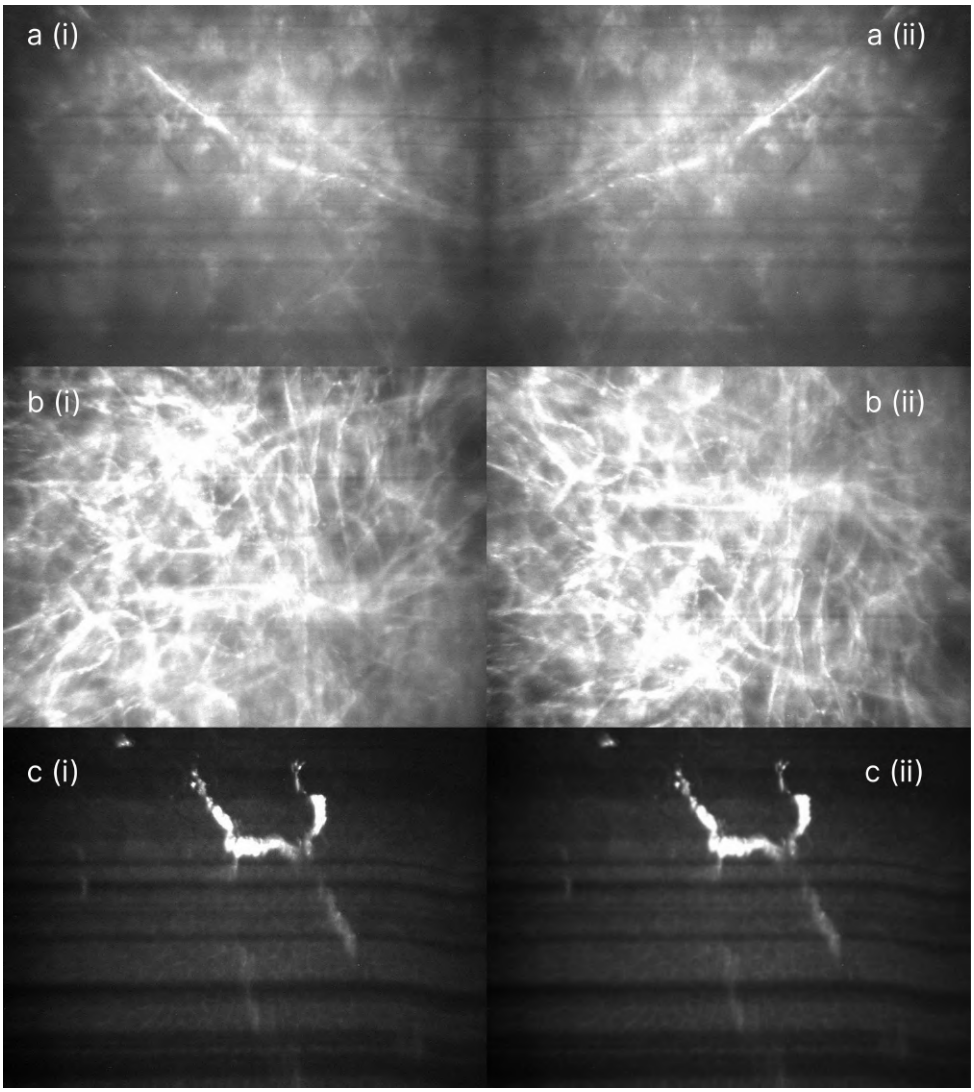


Figura 12 – Nas imagens acima, foram aplicadas as operações de *flip* horizontal a(ii), *flip* vertical b(ii) e *Gaussian blur* c(ii)

As imagens a(i) e b(i), presentes na Figura 12, correspondem a amostras originais da classe fúngica (FK), extraídas do conjunto de dados antes da aplicação das técnicas de aumento. A partir dessas amostras, foram geradas as imagens transformadas a(ii) e b(ii), por meio da aplicação das operações de espelhamento horizontal e vertical (*flip*), respectivamente. Essa estratégia tem como objetivo preservar as estruturas morfológicas

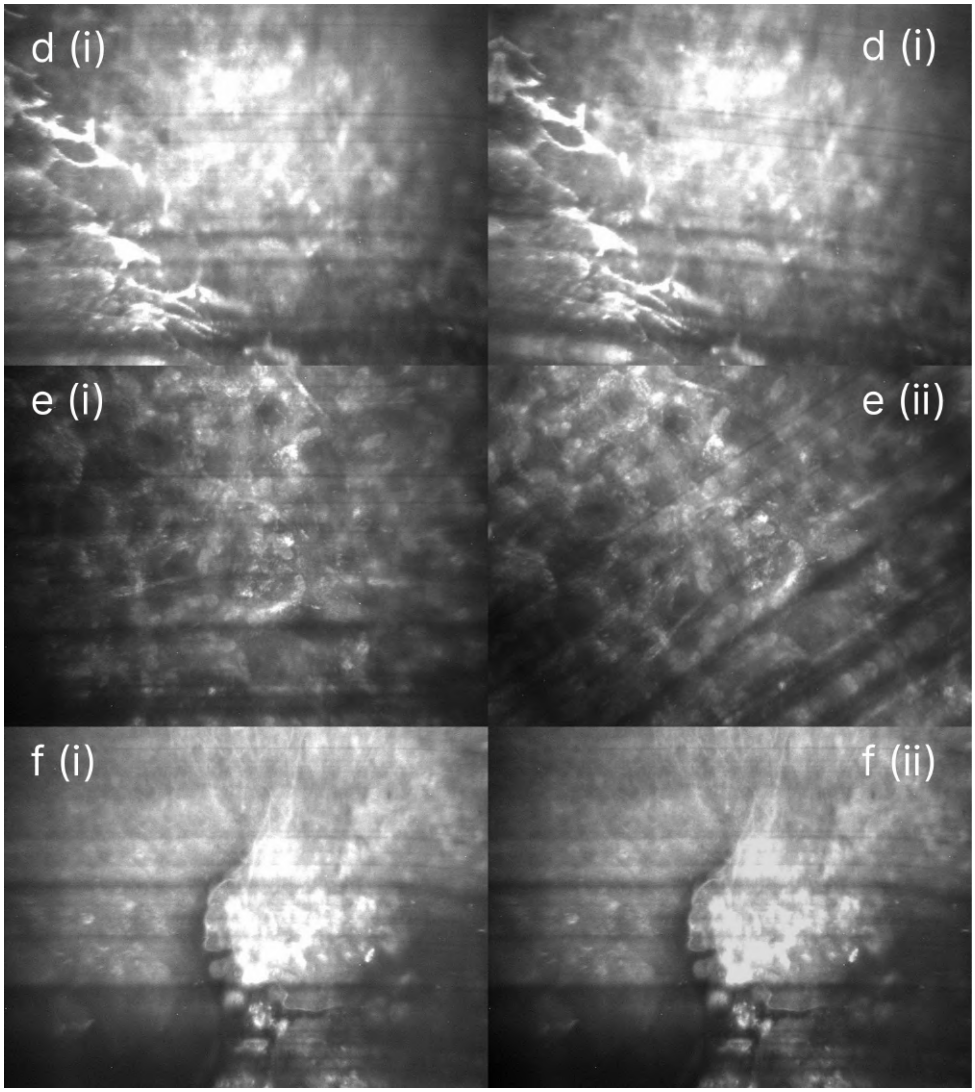


Figura 13 – Nas imagens acima, foram aplicadas as operações de rotação negativa d(ii), rotação positiva e(ii) e variação de cor (*color jitter*) f(ii)

enquanto introduz variações geométricas que auxiliam na capacidade de generalização do modelo durante o treinamento.

As imagens c(i), na Figura 12, e d(i), na Figura 13, representam amostras originais das classes Normal e Ceratite Não Especificada (NSK), respectivamente. A imagem c(ii) foi submetida a um leve desfoque gaussiano (*Gaussian blur*), com o propósito de suavizar ruídos presentes em imagens de microscopia. Já as imagens d(ii) e e(ii) passaram por rotações em sentidos opostos (negativo e positivo), empregadas para diversificar a orientação espacial das amostras.

Por fim, a subimagem f(ii), também na Figura 13, ilustra a aplicação da técnica de variação de cor (*color jitter*) a partir de uma amostra original da classe *Acanthamoeba* (AK). Essa transformação modifica aleatoriamente atributos visuais, como brilho, contraste e saturação, contribuindo para o aumento da robustez do modelo diante de diferentes condições de iluminação e variações de captura.

5.2 Treinamento e Validação Cruzada

A Etapa (III) corresponde à implementação e avaliação dos modelos estudados, por meio dos processos de treinamento e teste. O modelo possui um conjunto de hiperparâmetros ajustáveis, detalhados na Tabela 2, que incluem: taxa de aprendizado (*Learning rate*), número máximo de épocas (*Epochs*), tamanho do lote (*Batch size*) e critérios de parada antecipada (*Early stopping*), definidos a partir de um valor mínimo de melhoria (*Delta*) observado durante um intervalo de paciência (*Patience*). Além disso, as camadas convolucionais são congeladas por cinco épocas (*Freeze epochs*) para utilizar pesos extraídos dos treinamentos com o dataset ImageNet.

Tabela 2 – Hiperparâmetros utilizados.

Image Size	Batch size	K	Épocas	Learning rate	Paciência	Delta	Freeze Epochs
224×224	16	5	1000	0.0001	5	0.001	5

O treinamento foi conduzido utilizando um esquema de validação cruzada estratificada do tipo *k-fold*, com $k = 5$, garantindo que cada partição preservasse a proporcionalidade entre amostras reais e sintéticas em todas as classes. Em cada iteração, os modelos foram treinados com subconjuntos distintos dos dados e as métricas de desempenho foram agregadas ao final dos ciclos de validação.

As arquiteturas convolucionais pré-treinadas são empregadas como extratoras de características, organizadas em duas configurações híbridas principais: (i) MobileNetV3 combinada com VGG16 e (ii) DenseNet combinada com ResNet. Em ambas as configurações, apenas as camadas densas finais são inicializadas do zero, enquanto as camadas convolucionais pré-treinadas permanecem congeladas durante as cinco primeiras épocas de treinamento. Essa estratégia de *transfer learning* tem como objetivo estabilizar a fase inicial de otimização e permitir que as camadas totalmente conectadas se ajustem ao domínio específico da tarefa. Após esse período, todas as camadas são liberadas para *fine-tuning* supervisionado.

Para cada par de redes convolucionais, os mapas de ativação (*feature maps*) extraídos pela parte convolucional são reduzidos a vetores unidimensionais por meio de camadas lineares. Em todos os cenários, cada arquitetura produz um vetor de características de dimensão 512. Os dois vetores associados ao par de redes (MobileNetV3 e VGG16 ou DenseNet e ResNet) são concatenados e formam uma representação conjunta que é encaminhada a um bloco de atenção do tipo *Self-Attention* (SA). Esse módulo de atenção aprende a realçar componentes mais informativos da representação combinada, gerando um vetor de atenção que sintetiza as contribuições complementares de cada rede.

O vetor resultante do bloco de atenção SA é utilizado como entrada para a MLP final, responsável pela classificação. A MLP adotada é a mesma para todos os cenários

avaliados e é composta por uma camada de entrada compatível com a dimensionalidade do vetor de atenção, uma camada oculta com 2.048 unidades e uma camada de saída com quatro neurônios, correspondendo às classes do problema. A função de ativação empregada nas camadas internas é a *hardswish*, originalmente introduzida na família de modelos MobileNet.

5.3 Abordagens híbridas sem Bloco SA

Neste trabalho, o termo arquitetura híbrida refere-se a modelos de duas redes cujas representações são fundidas em nível de características por concatenação, formando um único vetor que alimenta o classificador. Assim, nas Abordagens I e II (Figura 14, a arquitetura é dita híbrida mesmo sem o bloco de self-attention, pois a hibridização decorre da combinação estrutural de dois fluxos convolucionais. O bloco SA, introduzido nas Abordagens III e IV (Figura 15, atua como um módulo de refinamento sobre a representação híbrida já concatenada.

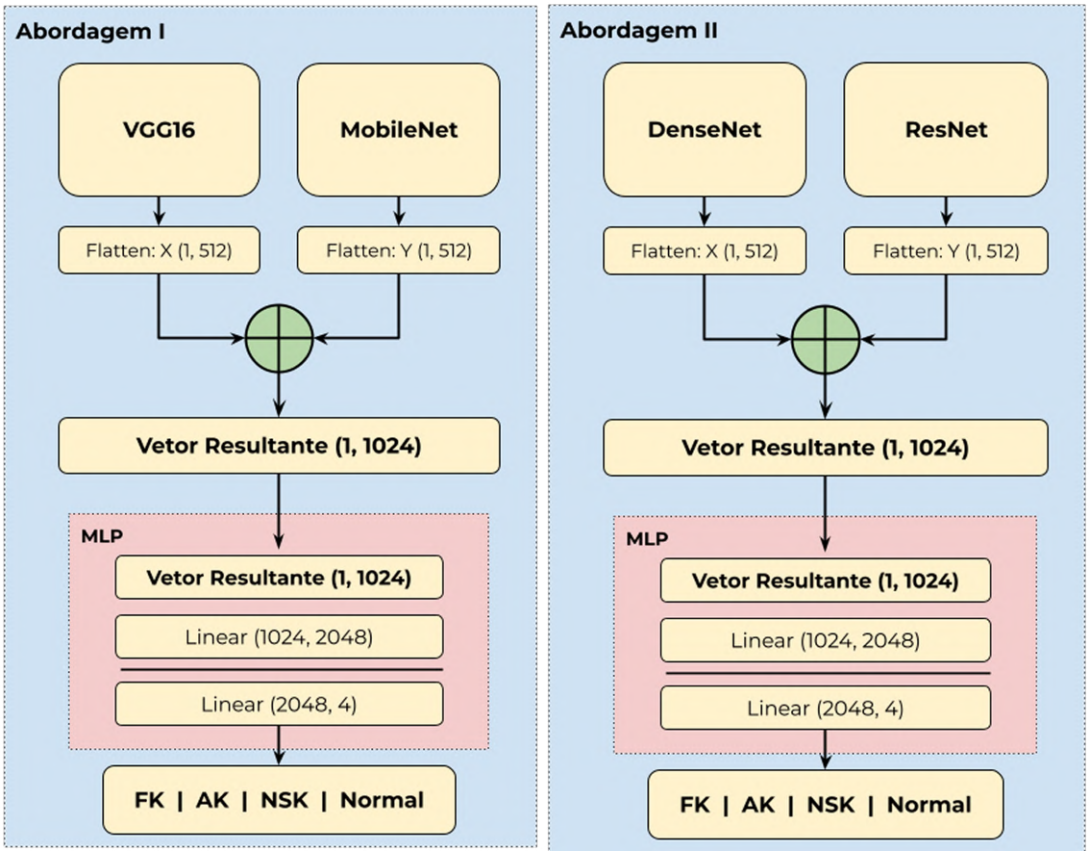


Figura 14 – Arquiteturas híbridas propostas sem módulo de atenção SA. À esquerda, o modelo utiliza os vetores de características das redes VGG16 e MobileNet; à direita, é a combinação das redes DenseNet e ResNet.

5.4 Abordagens híbridas com Bloco SA

Nesta etapa, foram desenvolvidas as Abordagens III e IV, que introduzem o mecanismo de SA na arquitetura híbrida. O objetivo desta inserção é permitir que o modelo pondere a importância relativa das características extraídas pelas diferentes redes base antes da classificação.

Conforme ilustrado na Figura 15, a Abordagem III combina as redes VGG16 e MobileNet, enquanto a Abordagem IV utiliza DenseNet e ResNet. Em ambos os casos, os vetores de características são achatados, concatenados e submetidos ao bloco SA, resultando em um vetor de dimensão $(1, 1024)$ que serve de entrada para a camada densa.

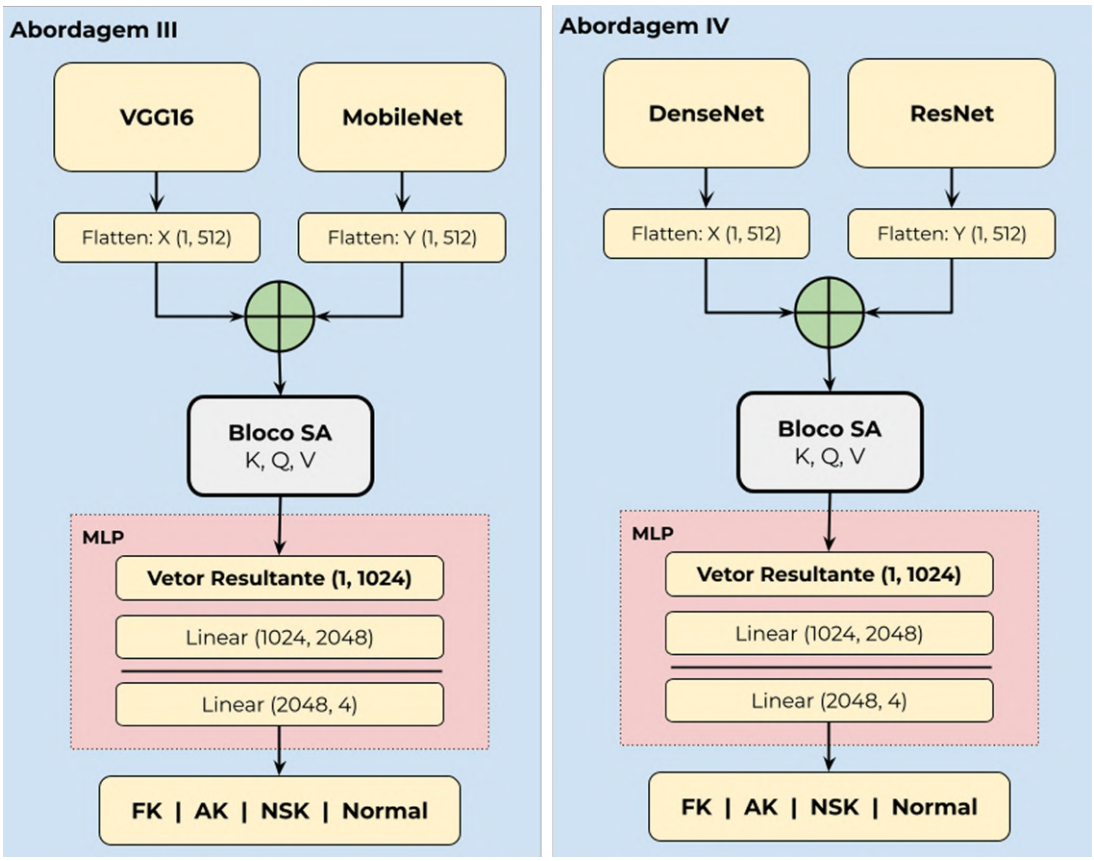


Figura 15 – Diagrama das arquiteturas propostas nas Abordagens III e IV. Ambas as configurações utilizam a fusão por concatenação de vetores de características extraídos de diferentes redes VGG16+MobileNet e DenseNet+ResNet, seguidos pelo mecanismo de *Self-Attention* para refinamento das features antes da classificação via MLP.

5.5 Avaliação

A Etapa IV (Figura 8) corresponde à fase de avaliação das abordagens propostas, realizada por meio de métricas quantitativas que mensuram o desempenho preditivo em tarefas supervisionadas. Neste trabalho, são utilizadas as seguintes métricas: acurácia, pre-

cisão (*Precision*), sensibilidade (*Recall*) e *F1-Score*. Cada métrica fornece uma perspectiva complementar sobre o comportamento do modelo, contribuindo para uma análise mais abrangente de sua performance.

Além disso, realiza-se uma comparação entre a abordagem com melhor desempenho e arquiteturas presentes na literatura. Essa análise inclui critérios de eficiência computacional, como o número de operações de ponto flutuante (FLOPs), o tamanho do modelo (em megabytes) e a quantidade total de parâmetros treináveis.

É feita uma análise comparativa entre a abordagem que mais se destacou neste projeto e o modelo utilizado por Essalat et al. (2023), visto que ambos os métodos foram aplicados na mesma base de imagens.

A etapa final do fluxo metodológico, representada como Etapa V na Figura 8, consiste na aplicação da técnica de interpretabilidade *Grad-CAM*. Esta análise é realizada especificamente sobre a arquitetura que apresentar o melhor desempenho nas métricas de avaliação, visando gerar mapas de ativação de classe. O objetivo é visualizar e validar as regiões da imagem que apresentam maior saliência e foram mais determinantes para a decisão do modelo.

5.6 Resultados e Discussão

A Tabela 3 apresenta os resultados médios de desempenho obtidos por cada abordagem avaliada, com base nas quatro métricas descritas nas Seções 4.5 e 5.5. As médias e os desvios padrão foram calculados a partir da validação cruzada com $K = 5$ *folds*, visando obter uma estimativa mais estável da capacidade de generalização dos modelos.

Tabela 3 – Métricas de desempenho obtidas nas diferentes abordagens avaliadas. As abordagens I e II não possuem mecanismo de atenção SA. Já as abordagens III e IV possuem o bloco de atenção e se destacaram em desempenho, indicando maior capacidade discriminativa do modelo.

	Média (%)			
Abordagem	**Acurácia**	**Precisão**	**Sensibilidade**	**F1**
VGG16	94,37 ± 2,27	90,89 ± 2,91	91,79 ± 1,11	91,01 ± 2,62
MobileNet	96,04 ± 0,56	93,77 ± 0,79	93,01 ± 1,63	93,30 ± 1,01
DenseNet161	92,48 ± 2,04	92,94 ± 1,89	92,48 ± 2,04	92,42 ± 2,07
ResNet101	85,69 ± 0,05	86,85 ± 0,82	85,69 ± 1,05	85,73 ± 1,14
(I) VGG16+MobileNet sem SA	83,79 ± 3,03	77,73 ± 3,14	78,29 ± 2,98	77,18 ± 2,75
(II) DenseNet161+ResNet101 sem SA	95,42 ± 0,47	92,54 ± 1,14	91,61 ± 1,22	91,93 ± 0,92
(III) VGG16+MobileNet com SA	**97,13 ± 0,86**	94,77 ± 1,28	95,80 ± 1,49	95,24 ± 1,34
(IV) DenseNet161+ResNet101 com SA	96,01 ± 1,02	**96,22 ± 0,84**	**96,01 ± 1,02**	**96,03 ± 1,00**

Os modelos VGG16, MobileNetV3, DenseNet161 e ResNet101 foram avaliados em suas arquiteturas originais, com o objetivo de estabelecer uma base de comparação para as abordagens híbridas propostas. Observa-se que a ResNet101 apresentou o menor

desempenho médio entre os modelos avaliados. Em contraste, a MobileNetV3 obteve desempenho superior aos demais baselines, mantendo alta acurácia e *F1-score*.

Adicionalmente, a Abordagem I apresentou uma queda expressiva de desempenho, sugerindo que a concatenação direta de características, sem um mecanismo de reponderação, pode introduzir redundância e dificultar a separação das classes. Por outro lado, as abordagens com SA (III e IV) apresentaram os melhores resultados médios, indicando que o módulo de atenção contribui para enfatizar características mais informativas antes da classificação. Uma hipótese para o desempenho da MobileNetV3, mesmo sem SA, é a presença de módulos do tipo *Squeeze-and-Excitation* em sua arquitetura (HOWARD et al., 2019).

Entre as abordagens VGG16+MobileNet (I) e DenseNet+ResNet (II), a Abordagem II apresentou melhor desempenho. Uma explicação plausível é que as conexões densas (DenseNet) e residuais (ResNet) favorecem o reaproveitamento de características e a propagação do gradiente, resultando em representações mais robustas para padrões texturais sutis presentes nas imagens.

Com o bloco SA, as abordagens III e IV alcançam médias semelhantes; entretanto, a Abordagem IV apresenta menor variabilidade (menor desvio padrão), sugerindo maior robustez à partição dos dados, além de maior sensibilidade média, o que é desejável para minimizar falsos negativos em imagens IVCM.

A Abordagem IV apresentou o melhor desempenho nas métricas relacionadas ao equilíbrio entre detecção e erros de classificação, atingindo os maiores valores médios de precisão (96,22%), sensibilidade (96,01%) e *F1-score* (96,03%). Embora não tenha obtido a maior acurácia média, esses resultados sugerem que a inserção do bloco de atenção SA após a concatenação das características extraídas pelas redes DenseNet161 e ResNet101 favorece uma decisão mais robusta, aspecto relevante no contexto de classificação de imagens IVCM.

O critério adotado para a seleção do melhor modelo baseou-se nas métricas de sensibilidade (*recall*) e *F1-score*, por refletirem, respectivamente, a capacidade do modelo de minimizar falsos negativos e de manter um compromisso entre sensibilidade e precisão. Esse critério é particularmente relevante no contexto do diagnóstico de ceratite infecciosa, em que a não detecção de casos patológicos pode acarretar atrasos na conduta clínica.

Considerando a abordagem IV, realiza-se uma comparação física entre suas características de desempenho e os modelos utilizados na literatura. A comparação se estende também a modelos de *deep learning* disseminados na comunidade científica. A Tabela 4 apresenta a disposição desses modelos, onde é possível compreender a posição da Abordagem IV em termos de desempenho preditivo e complexidade computacional, destacando seus pontos fortes frente a outros modelos de referência.

Tabela 4 – Tamanho físico (MB) e taxa de FLOPS da Abordagem IV frente a modelos de *deep learning*.

Modelo	**Parâmetros (M)**	**Tamanho (MB)**	**GFLOPs**	**MB/GFLOPs**	**M/GFLOPs**
Vision Transformer (ViT-L/16) (DOSOVITSKIY et al., 2021)	304.0	1161,00	61,55	18,86	4,94
VGG19 (SIMONYAN; ZISSERMAN, 2015)	143.7	548,00	19,60	27,96	7,33
VGG16 (LIU et al., 2020)	138.4	528,00	15,50	34,06	8,93
ConvNeXt-Base (LIU et al., 2022)	88.6	338,10	15,40	21,95	5,75
Abordagem IV	**78.0**	**297,49**	**15,75**	**18,88**	**4,95**
AlexNet (XU et al., 2021)	61.0	233,00	0,72	323,61	84,72
ResNet101 (LV et al., 2020; LINCKE et al., 2023)	44.5	170,50	7,80	21,86	5,71
DenseNet161 (ESSALAT et al., 2023)	28.7	110,40	7,70	14,34	3,73
Inception V3 (XU et al., 2021)	27.0	104,00	5,70	18,25	4,74
MobileNetV3-Large (HOWARD et al., 2017)	5.4	21,10	0,22	95,91	24,55

Com aproximadamente 78,0 milhões de parâmetros treináveis, a Abordagem IV possui o quinto maior número de parâmetros entre os modelos analisados. Em termos de complexidade computacional, a Abordagem IV requer cerca de 15,75 GFLOPs por imagem (Tabela 4), aproximadamente o dobro do custo das redes individuais DenseNet161 (7,70 GFLOPs) e ResNet101 (7,80 GFLOPs). Esse aumento era esperado devido ao uso combinado de duas arquiteturas e pode ser compensado pelas melhorias obtidas no desempenho de classificação.

Além disso, conforme apresentado na Tabela 4, a Abordagem IV apresentou uma razão favorável entre tamanho físico e custo computacional, com 18,88 MB por GFLOP. Observa-se também que o modelo apresenta uma baixa razão de parâmetros por GFLOP (4,95 M/GFLOPs), situando-se entre os menores valores do comparativo. Esses indicadores sugerem boa eficiência estrutural em termos de memória e custo aritmético, evidenciando a capacidade da abordagem proposta de atingir alto desempenho sem recorrer a arquiteturas com centenas de milhões de parâmetros.

A Tabela 5 apresenta uma comparação entre o DenseNet161, reportado como o melhor classificador em Essalat et al. (2023), e a Abordagem IV (DenseNet161 + ResNet101 com SA), selecionada como a mais eficaz neste estudo com base em sensibilidade e *F1-score*. Ressalta-se que a comparação é indireta, pois existem diferenças no conjunto de dados.

Tabela 5 – Resultados obtidos em Essalat et al. (2023) e pela Abordagem IV.

	Média (%)			
Modelo	Acurácia	Precisão	Sensibilidade	F1-Score
Densenet161 (ESSALAT et al., 2023)	96,93	92,52	94,77	93,53
Abordagem IV	**96,01**	**96,22**	**96,01**	**96,03**

Cabe destacar que, embora existam outros estudos relevantes na literatura sobre a classificação de imagens de ceratite infecciosa, não é possível realizar comparações diretas com esses trabalhos devido à indisponibilidade pública de suas bases de dados. O estudo de Essalat et al. (2023) representa, até o momento, a única iniciativa que disponibiliza abertamente um conjunto de imagens IVCM com rótulos patológicos validados, o que torna essa comparação relevante e viável.

Observa-se que a Abordagem IV supera o DenseNet161Essalat et al. (2023) em todas as métricas avaliadas. A acurácia da Abordagem IV atinge 96,01%, em comparação com os 96,93% do DenseNet161, denotando uma ligeira melhoria nas classificações corretas. Além da acurácia, há avanços nas métricas relacionadas ao desempenho por classe: a precisão aumenta de 92,52% para 96,22%, o *Recall* de 94,77% para 96,01% e o *F1-score* de 93,53% para 96,03%.

Tais resultados evidenciam que a Abordagem IV apresenta melhor equilíbrio entre sensibilidade e precisão, reduzindo simultaneamente a ocorrência de falsos negativos e falsos positivos, o que se reflete no aumento da taxa *F1*. Em conjunto com a validação cruzada, esse comportamento sugere maior estabilidade e melhor capacidade de generalização da arquitetura proposta.

Apesar do desempenho quantitativo, é fundamental compreender como a Abordagem IV toma suas decisões e quais regiões das imagens IVCM sustentam as predições. Para isso, empregou-se o Grad-CAM como técnica de interpretabilidade, permitindo gerar mapas de ativação que destacam as áreas da imagem com maior contribuição para a classe predita. Essa análise qualitativa possibilita verificar se o modelo se apoia em padrões morfológicos plausíveis e clinicamente relevantes, oferecendo evidências adicionais sobre a confiabilidade do processo de decisão.

Dois mapas de ativação Grad-CAM foram gerados, um para cada fluxo (DenseNet161 e ResNet101), a partir das respectivas camadas-alvo convolucionais selecionadas. Para a análise qualitativa, foram escolhidos exemplos em que a Abordagem IV classifica corretamente AK, FK e Normal, bem como exemplos de erro em que amostras de AK ou FK são preditas como NSK. Essa escolha se justifica porque a classe NSK apresentou o comportamento mais desafiador na validação cruzada, possivelmente devido à maior variabilidade intra-classe e à sobreposição de padrões visuais com as demais categorias.

A Figura 16 apresenta os mapas para um exemplo corretamente classificado como FK. No painel 16(a), referente à DenseNet161, a ativação é mais pontual, com múltiplos focos localizados, sugerindo a utilização de pistas visuais discretas e espacialmente delimitadas. Já no painel 16(b), referente à ResNet101, a ativação é mais abrangente e concentra-se predominantemente em uma única região próxima à borda da imagem, destacando partes de estruturas semelhantes a hifas, possível característica para FK. Em conjunto, os mapas evidenciam padrões de atenção distintos entre as redes e a contribuição de interpretação da ResNet101 se mostra maior quando comparada à DenseNet161.

A Figura 17 apresenta os mapas de ativação Grad-CAM para um exemplo corretamente classificado como NSK (ceratite não especificada). Essa classe tende a reunir padrões visuais heterogêneos, com estruturas diversificadas, e foi tratada como uma categoria específica por Essalat et al. (2023) com o objetivo de reduzir ambiguidades e melhorar a separação entre AK e FK.

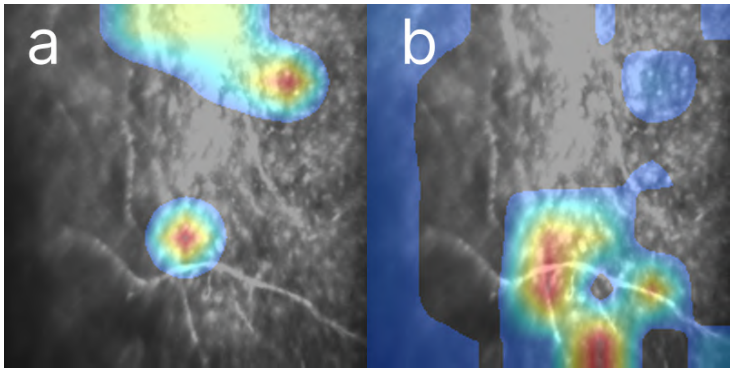


Figura 16 – Mapas de ativação Grad-CAM para um exemplo corretamente classificado como **FK** pela Abordagem IV. (a) DenseNet161: focos de ativação mais localizados e fragmentados. (b) ResNet101: ativação mais abrangente, com um foco dominante em uma região próxima a borda, destacando mais elementos presentes.

Observa-se que, no painel 17(a), referente à arquitetura DenseNet161, a ativação ocorre em múltiplas regiões, com focos localizados, sugerindo a utilização de diferentes pistas visuais para sustentar a predição. Já no painel 17(b), referente à arquitetura ResNet101, o mapa apresenta uma ativação mais contínua e concentrada em uma região dominante, indicando uma atenção mais integrada. Em conjunto, os mapas evidenciam padrões de atenção distintos e podem potencialmente somar-se entre as arquiteturas convolucionais. Além disso, apesar da presença de estruturas não relacionadas à classe (possíveis artefatos ou elementos do microambiente), essas regiões não apresentam ativações intensas, sugerindo que o modelo priorizou padrões mais discriminativos para NSK e foi menos influenciado por ruídos no campo de visão.

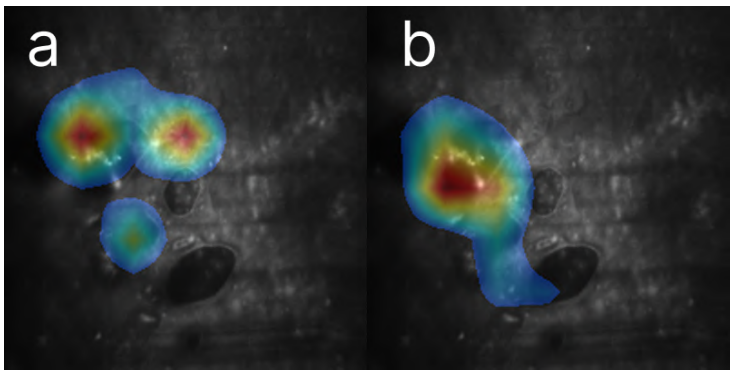


Figura 17 – Mapas de ativação Grad-CAM para um exemplo corretamente classificado como **NSK** pela Abordagem IV. (a) DenseNet161: múltiplos focos de ativação em regiões distintas. (b) ResNet101: ativação mais contínua e concentrada em uma região dominante.

É importante analisar também os casos de erro da Abordagem IV, pois ajudam a identificar padrões que levam a confusões entre classes. Em particular, a classe NSK reúne estruturas visualmente heterogêneas, o que pode aumentar a sobreposição de padrões com

outras categorias. A Figura 18 exemplifica uma predição incorreta, na qual o modelo prevê NSK, enquanto o rótulo verdadeiro é AK.

Observa-se que os Grad-CAMs dos das redes DenseNet161 (18a) e ResNet101 (18b) atribuem alta relevância a regiões de alto contraste e formato irregular próximas umas das outras. Esse comportamento sugere que, neste exemplo, o modelo priorizou padrões visuais compatíveis com NSK, reduzindo a sensibilidade a características discriminativas de AK, o que contribuiu para a confusão entre as classes.

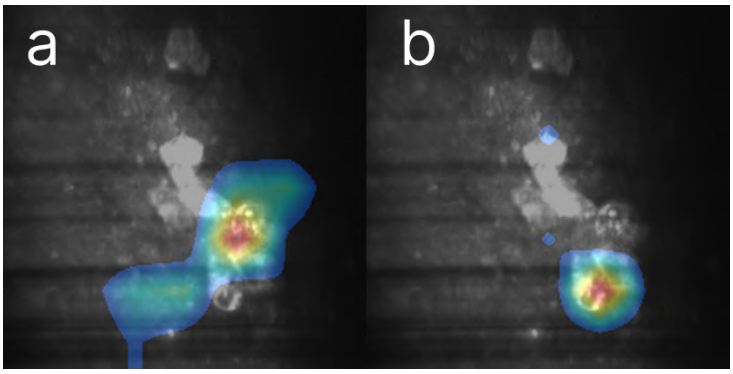


Figura 18 – Mapas de ativação de erro da Abordagem IV (real: **AK**, predito: **NSK**). Ambas destacam regiões de alto contraste e contornos irregulares como evidência principal para a decisão. (a) DenseNet161 apresentando uma ativação mais extensa e (b) ResNet101 com foco mais localizado.

6 Conclusão

Este capítulo sintetiza as principais contribuições deste trabalho no contexto da visão computacional aplicada à classificação de imagens de microscopia confocal *in vivo* (IVCM), com foco no diagnóstico assistido de ceratite infecciosa por meio da identificação automática de padrões compatíveis com a presença de agentes patológicos, incluindo fungos e protozoários do gênero *Acanthamoeba*.

A partir da análise do estado da arte (Capítulo 3), foram discutidas metodologias recentes baseadas em redes neurais convolucionais, com ênfase em estratégias de transferência de aprendizado e *fine-tuning*. Esse levantamento fundamentou a proposição de arquiteturas híbridas, motivadas pela hipótese de que a combinação de arquiteturas distintas pode produzir representações mais ricas e discriminativas para padrões sutis observados em imagens IVCM.

Entre os trabalhos relacionados, Liang et al. (2023) propõe uma abordagem híbrida baseada na concatenação de vetores de características extraídos por GoogLeNet e VGG16, seguida por um classificador MLP. Em comparação, a Abordagem IV desenvolvida neste estudo emprega DenseNet161 e ResNet101 como extratores de características e introduz um bloco de atenção do tipo SA após a concatenação, com o objetivo de reponderar de forma adaptativa as características mais informativas antes da classificação. Diferentemente de Liang et al. (2023), que não detalha a estrutura interna do classificador final, a Abordagem IV utiliza uma MLP com 2.048 neurônios e função de ativação *hardswish*, garantindo reprodutibilidade completa do arranjo proposto.

O estudo de Essalat et al. (2023) foi particularmente relevante por contextualizar a ceratite infecciosa do ponto de vista clínico e por disponibilizar publicamente a base IVCM-Keratitis, viabilizando a condução de experimentos reprodutíveis. Adicionalmente, a revisão sistemática apresentada por Kryszan et al. (2024) contribuiu para uma visão consolidada da produção científica na área, facilitando a identificação de tendências metodológicas e lacunas de investigação.

Os resultados obtidos indicam que a inserção do bloco de atenção SA em arquiteturas híbridas melhora o equilíbrio entre sensibilidade e *F1-score* quando comparada à mesma configuração sem o módulo de atenção, reforçando a utilidade do mecanismo de reponderação após a concatenação de características. No protocolo experimental adotado neste trabalho, a Abordagem IV apresentou desempenho superior ao DenseNet161 reportado em Essalat et al. (2023); entretanto, essa comparação deve ser interpretada com cautela, pois diferenças no conjunto de dados e no protocolo experimental inviabilizam uma equivalência estrita entre cenários.

Do ponto de vista de custo e armazenamento, a Abordagem IV apresentou razões favoráveis entre o tamanho do modelo e o custo aritmético (Tabela 4), sugerindo eficiência na relação memória-FLOPs quando comparada a arquiteturas mais pesadas. Ainda que o uso de duas redes base implique custo computacional superior ao de modelos individuais, os resultados evidenciam um compromisso competitivo entre desempenho preditivo e complexidade, especialmente no cenário em que métricas como sensibilidade e *F1-score* são prioritárias para minimizar erros clínicos relevantes.

Além do desempenho quantitativo, a aplicação do Grad-CAM possibilitou uma análise qualitativa do processo decisório do modelo, evidenciando que, em diferentes amostras, as arquiteturas base podem apresentar padrões de ativação complementares (ênfase em regiões distintas) ou parcialmente redundantes (ênfase em regiões semelhantes). Essa análise contribui para a interpretabilidade do modelo e auxilia na compreensão de como cada componente da arquitetura influencia a predição, oferecendo indícios sobre a plausibilidade visual dos padrões utilizados nas decisões.

Este trabalho atingiu os objetivos propostos ao investigar e validar abordagens híbridas para a classificação automática de imagens de ceratite infecciosa em IVCM, comparando a melhor configuração avaliada (Abordagem IV) tanto em termos de desempenho quanto em características físicas do modelo (tamanho e custo computacional). Como limitação, destaca-se que a comparação direta com a literatura foi possível de forma controlada apenas em relação a Essalat et al. (2023), uma vez que outros estudos não disponibilizam publicamente seus dados e protocolos, restringindo a reprodutibilidade e a comparação experimental em condições equivalentes.

Como trabalhos futuros, recomenda-se a validação externa em bases adicionais e cenários multivariados, visando avaliar a generalização frente a diferentes dispositivos e populações; (ii) estudos de ablação para quantificar a contribuição individual do bloco SA, da concatenação e das escolhas do classificador; e (iii) a integração de métricas e estratégias orientadas ao contexto clínico, como calibração de probabilidades e análise de erros por classe, de modo a ampliar a confiabilidade e a utilidade do sistema em apoio ao diagnóstico.

Referências

ADRIANO, M. et al. Perfil epidemiolÓgico dos casos de ceratite no brasil de 2018 a 2022. *Revista de Patologia do Tocantins*, Universidade Federal do Tocantins, v. 11, p. 336–339, 02 2024. Disponível em: <https://sistemas.uft.edu.br/periodicos/index.php/patologia/article/view/18817?utm_source=chatgpt.com>. Citado na página 1.

AUSTIN, A.; LIETMAN, T.; ROSE-NUSSBAUMER, J. Update on the management of infectious keratitis. *Ophthalmology*, v. 124, n. 11, p. 1678–1689, 2017. ISSN 0161-6420. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0161642016325295>. Citado na página 1.

CABRERA-AGUAS, M.; KHOO, P.; WATSON, S. L. Infectious keratitis: A review. *Clinical & Experimental Ophthalmology*, v. 50, n. 5, p. 543–562, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ceo.14113>. Citado 3 vezes nas páginas 1, 7 e 8.

COSSIO, M. Augmenting medical imaging: a comprehensive catalogue of 65 techniques for enhanced data analysis. *arXiv preprint arXiv:2303.01178*, 2023. Citado na página 27.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 248–255. Citado na página 11.

DOSOVITSKIY, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. Disponível em: <https://arxiv.org/abs/2010.11929>. Citado na página 42.

ERSAVAS, T.; SMITH, M. A.; MATTICK, J. S. Novel applications of convolutional neural networks in the age of transformers. *Scientific Reports*, Nature Publishing Group UK London, v. 14, n. 1, p. 10000, 2024. Citado na página 9.

ESSALAT, M. et al. Interpretable deep learning for diagnosis of fungal and acanthamoeba keratitis using in vivo confocal microscopy images. *Scientific Reports*, Nature Portfolio, v. 13, 06 2023. Citado 17 vezes nas páginas 9, 11, 1, 8, 19, 21, 22, 25, 26, 32, 33, 34, 40, 42, 43, 47 e 48.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020. Citado 2 vezes nas páginas 29 e 30.

HE, K. et al. *Deep Residual Learning for Image Recognition*. 2015. Disponível em: <https://arxiv.org/abs/1512.03385>. Citado 3 vezes nas páginas 9, 15 e 26.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 11.

HOWARD, A. et al. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019. Disponível em: <http://arxiv.org/abs/1905.02244>. Citado 5 vezes nas páginas 9, 13, 14, 26 e 41.

HOWARD, A. G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. Disponível em: <http://arxiv.org/abs/1704.04861>. Citado 3 vezes nas páginas 12, 13 e 42.

HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. Citado 3 vezes nas páginas 9, 14 e 26.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017. Citado na página 11.

KRYSZAN, K. et al. Artificial-intelligence-enhanced analysis of in vivo confocal microscopy in corneal diseases: A review. *Diagnostics*, v. 14, n. 7, 2024. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/14/7/694>. Citado 4 vezes nas páginas 2, 8, 19 e 47.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 9.

LIANG, S. et al. A structure-aware convolutional neural network for automatic diagnosis of fungal keratitis with in vivo confocal microscopy images. *Journal of Digital Imaging*, Springer, v. 36, n. 4, p. 1624–1632, 2023. Citado 4 vezes nas páginas 19, 21, 22 e 47.

LINCKE, A. et al. Ai-based decision-support system for diagnosing acanthamoeba keratitis using in vivo confocal microscopy images. *Translational vision science & technology*, The Association for Research in Vision and Ophthalmology, v. 12, n. 11, p. 29–29, 2023. Citado 5 vezes nas páginas 19, 20, 21, 22 e 42.

LIU, Z. et al. Automatic diagnosis of fungal keratitis using data augmentation and image fusion with deep convolutional neural network. *Computer methods and programs in biomedicine*, Elsevier, v. 187, p. 105019, 2020. Citado 3 vezes nas páginas 19, 22 e 42.

LIU, Z. et al. *A ConvNet for the 2020s*. 2022. Disponível em: <https://arxiv.org/abs/2201.03545>. Citado na página 42.

LV, J. et al. Deep learning-based automated diagnosis of fungal keratitis with in vivo confocal microscopy images. *Annals of translational medicine*, v. 8, n. 11, p. 706, 2020. Citado 4 vezes nas páginas 19, 20, 22 e 42.

PICCIALLI, F. et al. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, v. 66, p. 111–137, 2021. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253520303651>. Citado na página 7.

PRUSTY, S.; PATNAIK, S.; DASH, S. K. Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, Frontiers Media SA, v. 4, p. 972421, 2022. Citado na página 28.

SILVA, V. P. R. S. et al. DiagnÓstico e tratamento de ceratite infecciosa: Artigo de revisÃo. *Revista Contemporânea*, v. 5, n. 1, p. e7312, jan. 2025. Disponível em: <https://ojs.revistacontemporanea.com/ojs/index.php/home/article/view/7312>. Citado 2 vezes nas páginas 1 e 8.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Citado 3 vezes nas páginas 11, 12 e 26.

SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. Disponível em: <https://arxiv.org/abs/1409.1556>. Citado na página 42.

TAN, M. et al. Mnasnet: Platform-aware neural architecture search for mobile. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 2815–2823. Citado na página 13.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. Disponível em: <http://arxiv.org/abs/1905.11946>. Citado na página 12.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 16.

VAZQUEZ, F. J. B. Inteligência artificial aplicada à saúde: Qualidade na busca de diagnóstico. *Dataset Reports*, v. 3, n. 1, p. 93–100, set. 2024. Disponível em: <https://www.journals.royaldataset.com/dr/article/view/102>. Citado 2 vezes nas páginas 2 e 8.

VERMA, S. et al. Infectious keratitis: An update on role of epigenetics. *Frontiers in Immunology*, Volume 12 - 2021, 2021. ISSN 1664-3224. Disponível em: <https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.765890>. Citado 2 vezes nas páginas 7 e 8.

XU, F. et al. The clinical value of explainable deep learning for diagnosing fungal keratitis using in vivo confocal microscopy images. *Frontiers in Medicine*, Frontiers Media SA, v. 8, p. 797616, 2021. Citado 5 vezes nas páginas 19, 20, 22, 23 e 42.